TECHNISCHE UNIVERSITÄT MÜNCHEN
Fakultät für Informatik
Lehrstuhl für Bioinformatik

Phenotype Prediction From Evolutionary Sequence Covariation

Thomas A. Hopf

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des akademisches Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzende: Univ.-Prof. Gudrun J. Klinker, Ph.D.
Prüfer der Dissertation:

1. Univ.-Prof. Dr. Burkhard Rost
2. Assistant-Prof. Dr. Debora Marks
   (Harvard Medical School, Boston/USA)
3. Prof. Dr. Yitzhak Pilpel
   (Weizmann Institute of Science, Rehovot/Israel)

Die Dissertation wurde am 29.10.2015 bei der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 28.01.2016 angenommen.

Dedicated to my family

Abstract

Disentangling the relationship between genotype and phenotype is a major open question in biology. Since natural selection acts on phenotypes, the need to maintain organismal fitness leaves a trace of functional constraints in the underlying genotypes. We analyze abundantly available genotype data of homologous proteins from genomic sequencing for patterns of amino acid conservation and covariation. Based on statistical maximum entropy models of sequence coevolution to identify such evolutionary couplings, we developed computational methods to predict phenotypes that are difficult to obtain by experiment, on three different scales: (i) the three-dimensional structures of membrane proteins through coevolving sites in spatial contact; (ii) the interactions between pairs of proteins in protein complexes through coevolution of contacting sites in different molecules; and (iii) the quantitative, context-dependent impact of genotype changes on molecular and organismal function. For each of these approaches, we validated our phenotype predictions against available experimental data and show that, given sufficient genotype information, (i) accurate residue-residue contacts and three-dimensional structure predictions can be obtained for a wide range of proteins, allowing the de novo prediction of experimentally unsolved structures; (ii) coevolving positions between interacting proteins can be identified with high accuracy, enabling the detection of protein-protein interactions and the reconstruction of complex structures through molecular docking; and that (iii) the computed probabilistic, context-dependent effects of amino acid substitutions quantitatively agree with experimental measurements of biochemical function and organismal growth. In summary, our results show that statistical sequence modeling gives both a quantitative mapping from genotype to high-level phenotypes as well as biologically relevant intermediate molecular phenotypes such as protein structures or interactions. We anticipate that our approaches will allow to extract useful phenotypic information from the exponentially increasing amounts of sequence data and contribute to a deeper understanding of protein evolution and design as well as clinical applications.

Zusammenfassung

Die Entschlüsselung des Zusammenhangs zwischen Genotyp und Phänotyp ist eine ungelöste wichtige Fragestellung in der Biologie. Da die natürliche Selektion auf der Ebene des Phänotyps einwirkt, hinterlässt die Notwendigkeit, die Fitness des Organismus zu erhalten, Spuren funktioneller Beschränkungen auf der zugrunde liegenden Ebene des Genotyps. Wir analysieren die in großer Fülle verfügbaren Genotypdaten homologer Proteine aus Genomsequenzierungen hinsichtlich Mustern von Aminosäurekonservierung und -kovariation. Basierend auf statistischen Maximum-Entropie-Modellen der Sequenzkovariation zur Identifizierung dieser evolutionären Kopplungen entwickelten wir rechnergestützte Methoden zur Vorhersage von Phänotypen, die nur aufwändig mittels experimenteller Techniken gewonnen werden können, auf drei verschiedenen Ebenen: (i) Die dreidimensionale Struktur von Membranproteinen durch Koevolution von Positionen in räumlichem Kontakt; (ii) die Interaktion zwischen mehreren Proteinen durch die Koevolution von wechselwirkenden Positionen in Proteinkomplexen; sowie (iii) die quantitativen, kontextabhängigen Effekte von Veränderungen des Genotyps auf molekulare und organismische Funktion. Für jede dieser Methoden validierten wir unsere Phänotypvorhersagen gegen die verfügbaren experimentellen Daten und zeigen, dass, ausreichend Genotypinformation vorausgesetzt, (i) genaue Kontakte zwischen Aminosäureresten sowie dreidimensionale Strukturvorhersagen berechnet werden können, was die de novo-Vorhersage von experimentell ungelösten Proteinen ermöglicht; (ii) koevolvierende Positionen zwischen interagierenden Proteinen mit hoher Genauigkeit identifiziert werden können, was die Erkennung von Protein-Protein-Interaktionen sowie die Rekonstruktion von Komplexstrukturen mittels molekularem Docking erlaubt; und dass (iii) die probabilistische, kontextabhängige Modellierung von Aminosäuresubstitutionen experimentellen Messungen von biochemischer Funktion und organismischem Wachstum quantitativ entspricht. Zusammengefasst zeigen unsere Ergebnisse, dass die statistische Sequenzmodellierung sowohl eine quantitative Abbildung von Genotyp zu Phänotyp als auch biologisch relevante, dazwischenliegende molekulare Phänotypen wie Proteinstrukturen oder -interaktionen liefert. Wir erwarten, dass unsere Methoden die Extrahierung nützlicher phänotypischer Informationen aus der exponentiell zunehmenden Zahl an Sequenzen erlauben und zu einem tieferen Verständnis von Proteinevolution und -design sowie klinischen Anwendungen beitragen werden.

Publications

Manuscripts published during PhD studies:

Hopf, T. A. *et al.* Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife* **3** (2014)

Hopf, T. A. *et al.* Amino acid coevolution reveals three-dimensional structure and functional domains of insect odorant receptors. *Nat Commun* **6,** 6077 (2015)

Hopf, T. A. *et al.* Quantification of the effect of mutations using a global probability model of natural sequence variation. arXiv: 1510.04612 (2015)

Marks, D. S. *et al.* Protein structure prediction from sequence variation. *Nat. Biotechnol.* **30,** 1072–1080 (2012)

Kaján, L. *et al.* FreeContact: fast and free software for protein contact prediction from residue co-evolution. *BMC Bioinformatics* **15,** 85 (2014)

Tang, Y. *et al.* Protein structure determination by combining sparse NMR data with evolutionary couplings. *Nat. Methods* **12,** 751–754 (2015)

Sheridan, R. *et al.* EVfold.org: Evolutionary Couplings and Protein 3D Structure Prediction. *bioRxiv,* 021022 (2015)

Related manuscripts published prior to PhD studies:

Marks, D. S. *et al.* Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE* **6,** e28766 (2011)

Hopf, T. A. *et al.* Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* **149,** 1607–1621 (2012)

Other manuscripts (not related to the contents of this thesis):

Hopf, T. & Kramer, S. in *Discovery Science* (eds Pfahringer, B. *et al.*) *Lecture Notes in Computer Science* 6332, 311–325 (Springer Berlin Heidelberg, 2010)

Radivojac, P. *et al.* A large-scale evaluation of computational protein function prediction. *Nat. Methods* **10,** 221–227 (2013)

Hamp, T. *et al.* Homology-based inference sets the bar high for protein function prediction. *BMC Bioinformatics* **14 Suppl 3,** S7 (2013)

Acknowledgements

The past three years of research and writing this dissertation would not have been possible without the contributions of many other people. For this reason, I want to express my gratitude to:

Burkhard Rost, Chris Sander, and in particular Debora Marks, for making this transatlantic endeavor possible, exciting projects to work on, their mentorship, hospitality, inspiring discussions, and support in many other ways.

Gudrun Klinker, for chairing the examining committee, and Yitzhak Pilpel, for reviewing the dissertation.

All members of the Marks, Rost and Sander labs for an enriching scientific and personal environment; in particular Charlotta Schärfe for a great collaboration and proofreading of this dissertation, Frank Poelwijk and Ágnes Tóth-Petróczy for exciting discussions on protein evolution, Tobias Hamp for his input on protein complexes, John Ingraham for his contributions to the mutation effect project, Richard Stein for sharing his knowledge on maximum entropy models, and Li Yang Smith for her work on the transmembrane proteins.

My co-authors, for their valuable scientific contributions and insights.

The administrative and technical staff at TUM (Marlena Drabik, Inga Weise, Martina Hilla, Manuela Fischer and Tim Karl) and HMS (Maria Ferreira, Mason Miranda, Pamela Needham and Jake DeSilva, Research Computing), for their constant help and keeping things running smoothly.

The scientific Python community for providing such a fantastic set of tools, in particular the developers of the IPython/Jupyter notebook.

All my friends at home, in Munich and in Boston for a great time and offering distraction at times when things did not work as they should.

My family, for their never-ending and unconditional support throughout all the years.

Thank you all!

Contents

List of Figures

List of Tables

Abbreviations

3D	Three-dimensional/Three dimensions
APC	Average product correction
CASP	Critical Assessment of protein Structure Prediction
CDS	Coding sequence
CNS	Crystallography & NMR System
CPD	Compensated pathogenic deviation
DNA	Deoxyribonucleic acid
EC	Evolutionary coupling
EL	Extracellular loop
ENA	European Nucleotide Archive
FN	False negative
FP	False positive
GPCR	G-protein coupled receptor
GWAS	Genome-wide association study
HIV	Human immunodeficiency virus
HMM	Hidden Markov model
IL	Intracellular loop
MCC	Matthews correlation coefficient
NMR	Nuclear magnetic resonance
OR	Olfactory receptor
PDB	Protein Data Bank
RMSD	Root mean square deviation
RNA	Ribonucleic acid
RRM	RNA recognition motif
TMH	Transmembrane helix
TM score	Template modeling score
TN	True negative
TP	True positive

1. Introduction

Disentangling the relationship between *genotype* and *phenotype* is a fundamental challenge in biology and biomedicine[13,14]. Together with environmental factors, the genetic makeup of an organism (genotype) determines its observable characteristics (phenotype) on all scales of organization, including the molecular and cellular levels[15,16]. Genetic variation is the raw material for evolutionary processes through the fitness of the corresponding phenotypes; it influences the susceptibility of individuals and populations to disease and drug treatment and is therefore of key relevance to clinical applications[15,17,18]. Uncovering these intricate links between genotype and phenotype using dedicated experiments is however technically challenging, expensive and time-consuming. Today, the results of genomic sequencing efforts provide an abundant source of information on genotype data[17]. Can we mine this resource to extract the encoded phenotypes and predict the effects of genetic variation?

In the following, we will introduce the problem of mapping genotypes to phenotypes, show important underlying factors contributing to this relationship, and sketch how evolutionary genotype information and computational approaches can help to uncover phenotypic information.

1.1. The relationship of genotype and phenotype

Several fields of biology, including genetics and evolutionary biology, have long sought an understanding of how genotype determines phenotype, and of the intermediate links that connect these two layers of biological abstraction[15,16,18,19].

1.1.1. Biological relevance

A comprehensive understanding of genotype-phenotype relationships is particularly relevant to study the following three broad biological problems:

First, the characterization of evolutionary processes requires to assess the fitness (reproductive success) of different genotypes. Genetic variation is the raw material for evolution, natural selection however acts on the level of the corresponding phenotype and its interaction with the environment[16]. To connect genotypes and their fitness in an interpretable genotype-fitness map, we need to be able to determine the phenotypic consequences of mutational steps in genotype space[14,20].

Second, genetic variants have been associated with human disease in thousands of cases[17]. The identification of causal, disease-causing genotypes and the corresponding molecular phenotypes is a key challenge towards developing therapeutic interventions

and diagnosing individual disease susceptibility as well as drug efficacy in a clinical setting[16,18,21].

Third, how genotypes encode molecular phenotypes such as functional protein structures or interactions are central questions of molecular biology and biochemistry. Uncovering these links could help to learn about basic biophysical principles, give access to more phenotypic data through predictions, and aid the design of new biomolecular entities[16,22].

The precise instantiation of what constitutes a genotype and a phenotype is dependent on the particular application. In the most general definition, the genotype refers to the overall inherited genetic makeup of an individual organism, while the phenotype refers to its observable characteristics[15] that can be a complex set of attributes including molecular and morphological readouts[23]. Due to the vast space of possible genotypes and phenotypes, relative approaches that compare genotype differences to the corresponding phenotype differences may be more meaningful and feasible than an absolute, enumerative view of the genotype-phenotype map[15,16]. Possible differential views of the genotype space include e.g. wild-type genome sequences and variants with deletions of entire genes[24], or mutants of a certain protein-coding DNA sequence that result in single or higher-order amino acid changes to the wild-type protein. For the purposes of this work, we adapt a network-based view of genotype space, where nodes stand for individual protein sequences (which are directly determined by their genomic nucleotide sequence) and the edges between them correspond to single amino acid mutational steps that transform one sequence into another[16].

1.1.2. Experimental studies of the genotype-phenotype map

To uncover genotype-phenotype relationships, they have been characterized systematically in a variety of experimental screens that can be classified as either *forward* or *reverse genetics*[17] approaches. Forward genetics starts with a phenotype of interest (e.g. individuals affected by a certain disease) and tries to identify the genetic variants that cause the phenotypic difference. Modern experimental genotyping techniques, including high-throughput DNA sequencing, have made it possible to associate phenotypic differences with genetic variation on a whole-genome scale, as is evident from thousands of genome-wide association studies (GWAS) linking complex traits to genomic loci[15,17,18,25,26]. The identification of causal rather than associated variants however remains a challenge[18], as does the issue that only a small fraction of the observed heritability of diseases can be explained by the identified genetic variants[25].

Reverse genetics starts with the targeted alteration of a genotype of interest and attempts to identify differences in phenotype for each variant. Systematic gene deletion and RNA interference screens of entire genomes have been used to determine essential genes and gene function in different species[17,24,27–31]. More recently, deep mutational scanning experiments have allowed to comprehensively map the phenotypic consequences of mutational steps in sequence space around the wild-type se-

quences of DNA regulatory elements[32], catalytic RNAs[33] and protein molecules[34–71]. These experiments typically generate a multiplexed library of thousands of sequence variants which is then subjected to a functional assay that selects variants for a particular phenotype such as bacterial growth[71], protein stability[50] or ligand binding[44]. Mutational effects on the target phenotype are quantified by relating the frequency of each variant in the sequence library before and after functional selection, i.e. a ratio measuring variant enrichment or depletion[60]. The interpretation of the results of mutational scanning experiments strongly depends on the features of the used selection assay, including (i) the relevance of the target phenotype to the overall organism *in vivo* and (ii) the type and strength of applied selection pressure[49,59,62,71,72]. Despite these limitations, deep mutational scans provide a first systematic, empiric glimpse into the genotype-phenotype map for mutational variants of specific biomolecules[14,58,60].

A class of genotype-phenotype relationships that, like many others, has been historically studied in the framework of biochemistry rather than the abstract view of reverse genetics, is how genetic information in sequences gives rise to observable biological structures such as folded proteins, RNAs or protein interactions (in the following called *molecular phenotypes*)[16,73–75]. Since the determination of the first three-dimensional structures using X-ray crystallography[76] and nuclear magnetic resonance (NMR) experiments[77], coordinates for tens of thousands of proteins have been deposited in the Protein Data Bank (PDB)[78]. This enumerative genotype-phenotype map contains rich information about what structures particular sequences encode, how sequence variants causing amino acid substitutions change structure, and how different sequences encode similar structures. Nevertheless, three-dimensional structure information is missing for a large fraction of known protein-coding sequences because of experimental limitations and the low-throughput nature of structure determination experiments[79,80]. Although all information about the structure of a protein is in principle encoded in its sequence[81] (which is not always the case[22]), the mechanism how sequence (genotype) determines structure (phenotype) is still poorly understood and known as the *protein folding problem*[79]. Due to this missing link, experimental structure determination is still necessary on a per-protein basis for a large number of sequences.

1.1.3. Context-dependence of genetic variants

The analysis of genotype-phenotype relationships is complicated by the observation that the phenotypic consequences of genetic variants can depend on the genetic context, i.e. effects are modulated by the presence of variants in other loci to be stronger or weaker than expected[17]. This phenomenon is called *epistasis* and can affect variants in individuals of the same species as well as variation across different species[82].

The prevalence of epistasis in shaping phenotypic effects has been demonstrated in a number of studies using different approaches[39,63,83–94], but is still a topic of active debate[82,90,95–98]. Prominent examples for the importance of epistatic interactions include (i) disease-causing variants in humans that exist as wild-type allele in other

species (*compensated pathogenic deviations*, CPDs), but can be rescued by adjusting the genetic context of the human variant using a small number of substitutions observed in the other species[86–89]; (ii) amino acid co-dependencies determining the molecular specificity of protein-protein interactions[91–93]; and (iii) lower rates of evolutionarily observed amino acid substitutions than theoretically expected that suggest many variants are only acceptable in a certain sequence context[90,95].

Epistatic interactions play an important role in compensatory evolution, where the deleterious effect of one mutant in the wild-type background is set off by one or more additional substitutions[87,99,100]. The compensatory mutation may interact specifically with the deleterious mutation (*specific epistasis*), e.g. in a structural relationship in a protein structure[88,101], or act as a global enabler (*non-specific epistasis*) that compensates for deleterious effects independent of the particular substitution, e.g. by increasing protein stability to allow destabilizing mutations[16,17,102–104]. Such *permissive mutations* have been shown to be crucial to functional adaption[16,17,84,87,99,102,104–109]. A canonical example is the evolution of new substrate specificities in clinical isolates of TEM-1 β-lactamase to hydrolyze second- and third-generation cephalosporin antibiotics. Substitutions that rearrange active site residues to accommodate these non-natural ligands lead to a destabilization of the enzyme that is compensated by the globally stabilizing substitution M182T[105].

Although there is abundant evidence for compensatory mutations, the evolutionary processes by which they mainly occur remain unclear[100]. Since permissive mutations are neutral or advantageous, they may spread first and allow the occurrence of the otherwise deleterious and potentially function-changing mutation in a second step. Such cases have been described e.g. in the evolution of oseltamivir resistance in H1N1 influenza[84]. If both substitutions are deleterious on their own in the wild-type context, the compensatory substitution has to occur before being purged by purifying selection, as simultaneous occurrence of both variants seems unlikely[100].

A deeper theory of compensatory evolution will also require an understanding of the molecular mechanisms underlying the context-dependence of effects[17,82]. In proteins, complex interactions between amino acids determine their structures, functions and interactions with other molecules[16]. Not surprisingly, there are numerous reports of intra- and inter-molecular epistatic interactions between different loci in protein-coding sequences related to these and other features[14,16,44,63,91,93,100,103,110]. Since the biochemical phenotypes of protein molecules, including stability, have been shown to substantially influence the higher-level phenotypic effects and fitness of mutants[87,111], it seems plausible that the systematic consideration of amino acid interactions may be a key factor in accurately assessing variant effects and could yield detailed insights into the mechanistic emergence of epistasis.

1.2. Computational methods for phenotype prediction

The enormous size of genotype space and the combinatorial explosion created by epistatic interactions between different genetic variants present a major challenge to experimental studies of genotype-phenotype relationships. The systematic sequencing of the genomes of thousands of different species, as well as the genomes and exomes of human individuals, has created a deluge of genotype information without corresponding phenotypic information. The relevance and phenotypic consequences of hundreds of thousands of human genetic variants remain unknown[18]; current multiplexed techniques to assess the consequences of genetic variation in proteins are limited to the analysis of selected proteins and small isolated regions of sequence space (Section 1.1.2). Similarly, despite impressive advances in resolving the molecular details of protein structures and interactions, the sophisticated nature of these experiments precludes high-throughput applications to close the gap between the available amount of genotype information and three-dimensional structures[79]. In this setting, computational approaches to predict phenotype from genotype could be very useful[17,112]. Can we turn the problem around and leverage the information from myriads of mutation-selection experiments encoded in natural sequence variation, while considering epistatic interactions?

In the following, we will describe – for the case of protein-coding sequences – how evolutionary genotype information can contribute to the prediction of phenotypes, outline the challenges to obtain accurate predictions and review existing computational methods and their limitations.

1.2.1. Phenotype prediction from evolutionary sequence data

A possible predictive approach to the genotype-to-phenotype problem is to make use of evolutionary information. The observed sequences in genome and protein sequence databases are current viable endpoints of an ongoing evolutionary process, while many other genotypes were purged by purifying selection or have never been sampled by evolution (in combination with other factors such as neutral drift). Evolutionary sequences are therefore a snapshot of the positive outcomes of myriads of mutation-selection experiments testing the compatibility of certain genotypes with functional requirements on the phenotype level[16,18].

Assuming that a set of related genotypes, e.g. sequences belonging to the same protein family, is subject to the same or similar constraints, the selection for functional variants should leave a footprint of these constraints in the sequences in the form of patterns of conservation and acceptable variation[16]. Conservation usually refers to a lack of variation observed in a single locus (e.g., single-site conservation of a certain nucleotide or amino acid type). Epistatic interactions between different loci can however lead to higher-order patterns (*co-conservation* or *covariation*, e.g. amino acids in particular pairs of positions changing together) that are not visible from the

consideration of single sites alone. The identification of such covariation patterns could enable the detection of epistatic dependencies and allow their incorporation in prediction methods[16,82].

Conservation patterns in a set of sequences can then be used to predict the consequences of genetic variation by assessing how compatible the substituted variant is with evolutionarily observed variants (here: amino acid substitutions)[18]; the identified epistatic interactions could give information about three-dimensional structures of proteins and their complexes.

1.2.2. Functional effects of sequence variation

The accurate prediction of the effects of sequence variants is a key challenge in genetics and genomics regarding general studies of evolution and the assessment of individual disease risk[17,112,113] (Section 1.1.1). To characterize the consequences of sequence variation in nucleotide and amino acid sequences, a variety of computational methods has been developed[114–131]. These approaches are typically based on machine learning using a combination of input features such as evolutionary conservation and biochemical and structural information[18,132], with most predictive power usually originating from evolutionary conservation[120]. Most methods are trained in a supervised machine learning setting on existing experimental data to predict categorical outcomes such as deleterious vs. neutral effect; others explicitly incorporate three-dimensional structure information and force fields to quantify the consequences of mutations on protein stability[133–135].

Taking advantage of the fact that computational predictions can be easily calculated for a high number of variants, some of these methods have been used for the large-scale assessment of variant effects and the prediction of mutational landscapes[126,136–138]. It remains however unclear how reliable the performance assessment is for many of the machine learning approaches, and how well they generalize to new data. Recent work has demonstrated that the accuracy of some effect predictors critically depends on the data sets used for training and testing, and that variant data sets themselves may suffer from biased construction[113,132].

Surprisingly, despite ample evidence for the context-dependence of mutation effects (Section 1.1.3), the incorporation of the sequence context into prediction methods remains an exception[118,119,121] although results might benefit from its consideration[18,89]. Following the initial description of an explicit context-dependent model on the example of a small protein domain[139], epistatic models for variant prediction in HIV[140–142] and similar approaches using three-dimensional structure information[143–145] were described in parallel to this work but not systematically related to different functional features, experimental effects and improvements over non-epistatic models.

The consideration of epistatic interactions is closely related to the problem of choosing the right evolutionary depth for the input alignment, which is an aspect that is important but often not considered systematically[18,146]. The increased diversity of

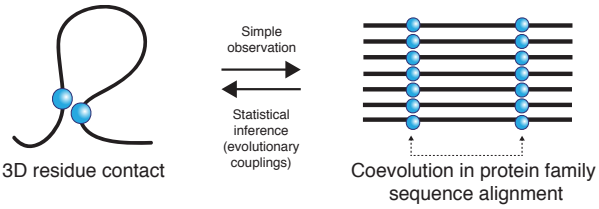

Figure 1.1.: Residue contacts leave a trace of amino acid coevolution. Evolutionary selection for folded protein structures leaves a trace of amino acid coevolution in protein families by maintaining favorable interactions between physically interacting residues. The inverse problem of inferring such evolutionary couplings, which would allow to predict 3D structure from sequence alignments, is statistically non-trivial due to transitive correlations and phylogenetic relationships, but can be solved using global statistical models. Adapted from Hopf *et al.*[2]

wide alignments can lead to an overestimation of acceptable substitutions due to the increased divergence of aligned sequences; reversely, narrow alignments may not capture all acceptable substitutions[18,127]. While epistatic models will not necessarily be able to account for the possible functional divergence of wider alignments, they could identify compensatory amino acid exchanges (e.g., compensated pathogenic deviations) and assess acceptable substitutions based on context rather than single site information alone. The use of epistatic models of sequence variation could therefore provide an opportunity to develop more accurate prediction methods and give insights into the co-dependencies governing the evolution of protein molecules.

1.2.3. Protein structures and interactions

Similar to the large number of uncharacterized functional effects of genetic variants, the scarcity of experimentally determined protein structures and their interactions creates a need to predict these molecular phenotypes computationally from sequence information[79]. Over the last decades, a large number of different methods addressing this problem have been developed which can be coarsely classified as either comparative modeling, force field-based *de novo* or machine learning approaches[4]. Comparative modeling exploits the observation that the three-dimensional structures of proteins remain similar even as their primary sequences diverge by transferring coordinates from a solved protein to a sequence-similar modeling target[147,148]. If a sufficiently similar template can be identified, comparative modeling tends to give the most accurate solutions compared to other strategies[149] and can also be applied for protein complexes[150].

In the absence of identifiable template structures, *de novo* methods based on assembling sequence-similar structural fragments or protein subunits using molecular force fields have emerged as a viable alternative for small protein molecules (<100 residues)[151], but are challenged by the vast size of conformational search space and the quality of current empirical force fields[79]. A different route is the use of machine learning-based methods to gain information about structural features from sequence,

their limited predictive accuracy however presents an obstacle to the analysis of individual proteins of interest[152,153]. To overcome the limitations of these purely computational *de novo* methods, hybrid data-driven approaches have been developed that integrate additional data, e.g. from NMR experiments, to bias the space of possible solutions towards the correct answer[154,155].

An orthogonal way of addressing the structure and interaction prediction problem is by using evolutionary information. Different from mutation effect prediction, early observations of compensatory evolution in protein sequences[156,157] led to the development of methods exploiting epistatic dependencies within[158–162] and between proteins[163–165]. These *correlated mutation* approaches are based on the hypothesis that the coevolution of amino acid residues in spatial contact is necessary to maintain functional protein structures and complexes (Figure 1.1). Despite their conceptual appeal, these early works found that residue pairs with a strong coevolution signal correspond to structural contacts only in relatively few cases; subsequent investigations suggested that the observed coevolution signal derives from allosteric functional couplings rather than structural contacts[166–168].

However, it was shown that the lack of identified structural contacts was the consequence of phylogeny and transitive effects (where direct coevolution of contacting residues causes strong, indirect correlations) that could be partly remedied by the use of global probability models instead of analyzing pairs of positions independently of each other[4,139,169–172]. The development and application of such global approaches based on the maximum entropy principle[8,139,173,174], a related partial correlation formalism[175,176] or Bayesian network modeling[177,178] demonstrated that residue-residue contacts in protein domains and a selected protein interaction could be accurately identified from sequences alone. Later methodological developments further improved the accuracy of the high-scoring pairs as predictors of residue contacts[179–182].

The identification of epistatic pair couplings corresponding to structural contacts enabled the *de novo* prediction of protein three-dimensional structures even for large molecules far outside the reach of previous methods[8,9,183–188], their conformational changes[9,189,190] and of functionally important residues[9]. Anecdotal evidence also suggested that the structural details of protein-proteins interactions can be elucidated from sequence covariation[9,173,177,191].

Despite strong requirements on the number and diversity of available protein sequences, coevolution approaches provide a chance to obtain detailed structural information for unsolved proteins of biological interest for the first time[4,172]. The initial results on protein complexes[9,173,177,191] additionally suggest that residue interactions between proteins could be characterized at a more general scale by further developing these preliminary prediction approaches to cover arbitrary protein interactions.

1.3. Thesis contributions

In this work, we address the problem of predicting phenotypes from protein-coding sequences to close the gap between abundant genotype information and scarce phenotype data. We apply probabilistic maximum entropy modeling of protein families to identify epistatic interactions between residues from evolutionary covariation patterns and develop approaches to predict phenotypes on different biological scales: (i) the three-dimensional structures of unsolved transmembrane proteins of interest, including insect olfactory receptors; (ii) protein-protein interactions and their structural details; and (iii) the quantitative, context-dependent effects of mutations transforming one sequence into another (Chapter 2). We systematically evaluate our prediction methods against experimental phenotype data, demonstrate that accurate predictions can be obtained for all three target phenotypes if there is sufficient evolutionary information, and provide *de novo* phenotype predictions for experimentally uncharacterized candidate proteins (Chapter 3). We conclude with a discussion of the implications of our and related work as well as of the complex interplay between predicted and experimental phenotypes on different scales, and outline areas for future development (Section 3.4).

2. Methods

This chapter describes the details of our methods to predict protein phenotypes from evolutionary sequence variation and how they were evaluated against existing experimental data. The main prediction steps include the generation of a protein family sequence alignment, the inference of a statistical model of sequences from this evolutionary record, and the subsequent prediction of phenotypes using the statistical model and its parameters.[i]

2.1. Alignments of evolutionary sequences

The input for statistical coevolution analysis is a multiple sequence alignment of homologous protein sequences. The following section outlines how homologs of a protein of interest were identified and aligned, as well as additional processing steps that had to be applied for protein-protein interactions.

2.1.1. Sequence searches using HMM-based methods

For all proteins of interest, multiple sequence alignments of the protein family were generated using iterative hidden Markov model (HMM)-based sequence similarity search tools[192,193]. These tools display greater accuracy and sensitivity in identifying and aligning more distant homologs than earlier methods[193–195] by modeling position-specific amino acid preferences and deletion and insertion probabilities[196].

Given the highly divergent nature of many protein families[197], a key challenge is to generate alignments of appropriate evolutionary depth, which are neither too narrow nor too wide (Section 1.2.2). If an alignment is too narrow, i.e. contains only sequences similar to the query sequence, it might not contain enough variation to detect patterns of context-dependence or provide enough samples for statistical inference. If an alignment is too wide and contains sequences that diverged too far from the query sequence, the assumption of conserved function (isofunctionality or isostructurality) may be violated[198].

To control for the evolutionary depth of alignments, we used a length-dependent bit score homolog inclusion threshold rather than an E-value threshold. Whereas E-values as a measure of statistical expectation are dependent both on the length of the query protein and the size of the sequence database, bit scores directly measure sequence similarity by scoring the agreement of site amino acid distributions[199]. Setting the

[i] This chapter unifies and extends the methods introduced in references[1–4,9]

same expectation for the average sequence similarity per residue (bits/residue) allows to obtain sequences of comparable evolutionary divergence across different proteins.

For this work, we used the *jackhmmer* application from the HMMER software suite[192] with 5 search iterations, a domain-specific bit score inclusion threshold of $x * L$ (parameter: --incdomT $x * L$, where x=expected similarity in bits/residue and L=length of query sequence region) and otherwise default settings. As a proxy for choosing a suitable evolutionary depth, we used our previously established strategy that trades off between finding as many homologs as possible while retaining alignment coverage for most positions of the query sequence[9]. We found that a threshold of 0.5 bits/residue was a robust starting point for most applications and then systematically decreased or increased the threshold to generate wider or narrower alignments. The exact thresholds used for all analyzed proteins are described in detail in the appended publications.

While not explored further here, we note that alignments of improved quality can in principle be obtained by separating the steps of homolog detection and alignment. In this case, after performing HMM-based searches, the identified set of sequences can be realigned using multiple alignment software that scales to tens or hundreds of thousands of sequences, such as Clustal Omega[200].

2.1.2. Choice of input sequence database

Depending on the particular phenotype to be predicted, different input protein sequence databases are better suited to the task. When predicting phenotypes for single proteins, usually only non-redundant sequence information is relevant, but not the particular identities of the sequences. In these cases, we gathered homologs from the pre-clustered UniRef100 database[201], which merges identical sequences and subfragments in UniProt[202] into maximally informative cluster representatives. Using this compressed sequence set reduces computational costs during alignments and statistical inference by removing non-informative samples. However, in cases where the particular identities of the proteins were relevant, such as predicting interactions between two proteins, we retained all information and used the full UniProt database[202]. We also created custom sequence databases for a protein family of interest (insect olfactory receptors), as many of these sequences had not yet beet deposited in public databases.

2.1.3. Sequence matching for protein complexes

The analysis of patterns of sequence covariation between two or more proteins in a heterocomplex requires to match the interacting partners in each species, as specific correlated residue exchanges are usually only to be expected between proteins that actually bind. The generation of complex alignments is therefore a two-step process, in which homologs for each of the interacting proteins A and B (monomers) are first

identified independently as detailed in Section 2.1.1 and then matched as interaction partners.

Generally, no comprehensive information will be available which homologs A′ of protein A interact with which homologs B′ of protein B, i.e. the conservation and specificity of the protein interaction across orthologs and paralogs is unknown (in this work, we only consider protein interactions within one species, but not between species, e.g. host-virus interactions). We therefore devised two matching strategies that use proxies which homologs are most likely to interact.

Sequence matching based on genomic distance

The first strategy is based on the assumption that in bacteria and archaea, proteins that are on the same operon are more likely to interact than proteins more distant on the genome[93,173]. To match sequences based on genomic distance, we retrieved genomic location information for the coding sequences (CDS) of all identified homologs A′ and B′ from the ENA database[203] and paired a particular A′ and B′ if the following three conditions were met:

1. The CDSs of A′ and B′ are present together on at least one ENA sequence entity (e.g. whole genome sequence or sequence contig).
2. Of all such possible pairings in a species, A′ and B′ are mutually closest to each other (i.e., there is no B″ which is closer to A′ than B′, and no A″ which is closer to B′ than A′). The distance between A′ and B′ is measured as the number of nucleotides between the two CDSs.
3. We additionally imposed a maximum distance threshold of 10000 nucleotides on any matched pair, to remove pairs which are mutually closest but distant on the genome, and therefore less likely to interact.

While this strategy is not applicable to eukaryotic protein complexes, it allows to match more than one pair A′-B′ per species, if the complex is present in multiple copies and these are co-located on the genome (e.g. membrane transporters of different substrate specificities). The one to one pairing of sequences applied here does not take into account possible cross-interactions, i.e. that A′ might interact non-exclusively with multiple B′ and vice versa.

Sequence matching based on closest homolog per species

The second strategy is based on the assumption that the interaction between proteins A and B is most likely to be conserved between their orthologs in other species. As a simple proxy for more elaborate ortholog identification approaches, for each species, we identified the sequences A′ and B′ that have the highest sequence identity to A and

B and paired these while discarding all other sequences A″ and B″. This strategy can also be applied to eukaryotic proteins, at the expense that only one pair per species can be matched and less sequence information is available for statistical inference.

Independent of the matching strategy applied, the aligned sequences for all matched pairs A′-B′ were extracted from the respective monomer sequence alignments and concatenated into one joint sequence each, yielding a concatenated sequence alignment for the protein complex. This alignment was used for statistical inference in the same way as monomer protein alignments.

2.1.4. Alignment post-processing

To ensure only high-confidence information about amino acid constraints was used during statistical coevolution inference, a post-processing step was applied to all alignments. Positions in the analyzed target sequence containing gaps instead of amino acid symbols in too many of the aligned sequences were excluded from analysis (>50% or >80% gaps). Similarly, sequence fragments that only aligned to a limited part of the target sequence were removed from the alignment (<50% of target sequence residues covered). The exact filters applied for each application are detailed in the appended publications.

2.2. Statistical modeling of evolutionary sequences

From the evolutionary record contained in the multiple sequence alignment, we identified evolutionary constraints on the sequences by inferring a global statistical model that captures patterns of amino acid conservation and covariation. The following section is mainly based on the work by Ekeberg *et al.*[180] and related work by Balakrishnan *et al.*[181].

2.2.1. Pairwise maximum entropy model of sequences

To identify single-site and higher-order amino acid constraints for a protein family, we inferred the parameters of a maximum entropy model (undirected graphical model, Potts model) over protein sequence space that explains observed patterns in the data using hidden, direct constraints[8,139,173,174,179–182]. The use of such a probabilistic approach is motivated by the observation that due to the cooperative nature of protein molecules, simple measures of positional co-dependency like mutual information are impaired by transitive effects masking direct constraints[139,169,170]. Of all models over discrete sequences, the maximum entropy model is the least biased distribution consistent with the single-site and higher-order frequencies of amino acids in the sequence alignment[139]. In this work, we used a model consistent with single-site and pair frequencies to limit the number of model parameters to $\mathcal{O}(N^2)$, but note that models of higher order are theoretically possible given large enough protein families[174].

Under the pairwise model, the probability of any amino acid sequence $\sigma = (\sigma_1, \ldots, \sigma_N)$ of length N is defined as

$$P(\sigma) = \frac{1}{Z} \exp\left(\sum_{i=1}^{N} \boldsymbol{h}_i(\sigma_i) + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \boldsymbol{J}_{ij}(\sigma_i, \sigma_j) \right). \tag{2.1}$$

The external fields $\boldsymbol{h}_i$ (conservation) and pair couplings $\boldsymbol{J}_{ij}$ (covariation) describe the family- and site-specific constraints on the likelihood of amino acid assignments σ_i and σ_j at sites i and j. Each variable σ_i can be one of the 20 naturally occurring amino acids or the gap character encoding deletions relative to the target sequence. The *partition function* Z is defined as

$$Z = \sum_{\sigma} \exp\left(\sum_{i=1}^{N} \boldsymbol{h}_i(\sigma_i) + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \boldsymbol{J}_{ij}(\sigma_i, \sigma_j) \right) \tag{2.2}$$

and sums over all possible 21^N system configurations $\sigma = (\sigma_1, \ldots, \sigma_N)$ to ensure that $P(\sigma)$ is a valid probability distribution ($\sum_\sigma P(\sigma) = 1$). Due to the exponential number of summations, calculating Z is infeasible for our applications as most protein sequences analyzed in this work have $N > 50$.

Finding the evolutionary constraints now consists of solving the inverse problem of inferring the model parameters from the sequence data.

2.2.2. Model inference from sequence data

The parameters of the statistical model could in principle be inferred directly from the sequence data in the alignment using standard maximum likelihood estimation. This approach is however hindered by three issues: (i) The intractability of Z prohibits the calculation of the likelihood function; (ii) the phylogenetic relationships between sequence samples in the alignment violate the assumption of independently drawn samples; and (iii) the number of parameters in the model exceeds the number of sequence samples by orders of magnitude, making parameter estimation susceptible to overfitting. Approximate solutions to address these issues in the statistical inference process are given in the following sections.

Pseudo-likelihood approximation

A standard way of inferring the parameters of statistical models is maximum likelihood estimation, which chooses the set of parameters that maximizes the probability of the observed data samples. For the pairwise model introduced in Section 2.2.1 and

a set Σ of aligned sequences σ, the likelihood function is defined as

$$\begin{aligned}\mathcal{L}(\boldsymbol{h}, \boldsymbol{J}) = P(\Sigma \mid \boldsymbol{h}, \boldsymbol{J}) &= \prod_{\sigma \in \Sigma} P(\boldsymbol{\sigma} \mid \boldsymbol{h}, \boldsymbol{J}) \\ &= \prod_{\sigma \in \Sigma} \frac{1}{Z(\boldsymbol{h}, \boldsymbol{J})} \exp\left(\sum_{i=1}^{N} \boldsymbol{h}_i(\sigma_i) + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \boldsymbol{J}_{ij}(\sigma_i, \sigma_j)\right)\end{aligned} \tag{2.3}$$

and is a function of the model parameters $\boldsymbol{h}$ and $\boldsymbol{J}$. Since the likelihood function and its partial derivatives depend on the intractable calculation of $Z(\boldsymbol{h}, \boldsymbol{J})$, maximum likelihood estimation is not applicable to our problem[180]. For tractable parameter estimation, several previously established approaches have been applied and evaluated in the context of pairwise graphical models of protein sequences, including gradient ascent with Monte Carlo sampling[139], message passing[173], and mean-field[8,174] or pseudo-likelihood approximations[179–181,204].

Here, we use the pseudo-likelihood approach, which approximates the full likelihood for each sequence $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_N)$ by a product of conditional likelihoods for each site i, i.e.

$$P(\sigma_1, \ldots, \sigma_N | \boldsymbol{h}, \boldsymbol{J}) \approx \prod_{i=1}^{N} P(\sigma_i \mid \boldsymbol{\sigma} \setminus \sigma_i, \boldsymbol{h}, \boldsymbol{J}). \tag{2.4}$$

By conditioning the probability of observing a certain amino acid σ_i in site i on the rest of the sequence ($\boldsymbol{\sigma} \setminus \sigma_i$), the global partition function $Z(\boldsymbol{h}, \boldsymbol{J})$ cancels out, leaving a tractable local normalization over all possible 21 amino acids a at site i:

$$P(\sigma_i \mid \boldsymbol{\sigma} \setminus \sigma_i, \boldsymbol{h}, \boldsymbol{J}) = \frac{\exp\left(\boldsymbol{h}_i(\sigma_i) + \sum_{j \neq i} \boldsymbol{J}_{ij}(\sigma_i, \sigma_j)\right)}{\sum_a \exp\left(\boldsymbol{h}_i(a) + \sum_{j \neq i} \boldsymbol{J}_{ij}(a, \sigma_j)\right)} \tag{2.5}$$

Through the above factorization, the time complexity of calculating the likelihood of the data therefore gets reduced from $\mathcal{O}(21^N)$ to $\mathcal{O}(|\Sigma| \cdot N^2)$ for the pseudo-likelihood, and parameters can be inferred from the data using standard iterative optimization procedures[180]. The pseudo-likelihood approximation has been shown to yield parameter estimates similar to the full likelihood solution in simulations[179] and outperforms all other inference methods tested in protein contact prediction applications[180]. In the theoretical limit of infinite data, the pseudo-likelihood estimate will converge to the true parameters if the inferred model is the true underlying distribution (consistent estimator)[179].

Sample reweighting

A central assumption of maximum likelihood inference is that the data samples are independent and identically distributed (*i.i.d.*). Biological sequences violate the independence assumption on several levels as (i) they are related by phylogeny, i.e. they

originate from common ancestors, (ii) the sequences deposited in databases are subject to which species have been sequenced so far and (iii) evolution usually has not explored the full space of possible functional sequences[197]. Functional divergence of the sequences in a protein family and different selective pressures can also lead to a violation of the identical distribution assumption, although there is no *a priori* reason to believe that protein sequences are distributed exactly according to the inferred model.

To address the dependence of sequences in an alignment, in particular points (i) and (ii), we followed a previously established strategy to reduce the influence of densely sampled parts of sequence space by reweighting samples based on their similarity to each other[8,173,174,180]. In this strategy, each sequence σ is assigned a weight

$$w(\sigma) = \frac{1}{m(\sigma)} \tag{2.6}$$

where

$$m(\sigma) = \left|\left\{\sigma' \mid \sigma' \in \Sigma \wedge \text{seqid}(\sigma, \sigma') \geq xN\right\}\right| \tag{2.7}$$

measures the number of sequences σ' in the alignment Σ that have more than a fraction of $x \in (0, 1]$ identical residues (seqid) to sequence σ with length N. Intuitively, this approach assigns a joint weight of 1 to a cluster of sequences that are all similar to each other above a threshold of x, thereby treating those sequences as one effective sample. For most of our applications, we used a similarity threshold of $x = 0.8$ to cluster sequences at 80% sequence identity.

The sequence weights $w(\sigma)$ were then used to multiplicatively adjust the contribution of each sequence to the log-transformed pseudo-likelihood function. The redundancy-reduced *effective number of sequences* in the alignment is measured by

$$M_{\text{eff}} = \sum_{\sigma \in \Sigma} w(\sigma) \tag{2.8}$$

and sums over the reweighted contribution of sequences σ in the alignment Σ. Here, M_{eff} is used to assess the amount of available sequence information when choosing the evolutionary depth of protein family alignments.

We note that besides the simple strategy used here, more elaborate and effective approaches to address the phylogenetic relationships between samples could in principle be applied[205].

Regularization

The number of parameters of the pairwise maximum entropy model (Equation 2.1, $\binom{N}{2} \cdot (q-1)^2 + N \cdot (q-1)$ free parameters for a protein of length N and $q = 21$ states) usually exceeds the available number of samples by several orders of magnitude, making parameter inference highly susceptible to overfitting. E.g., a protein of length

$N = 100$ has approximately $2 \cdot 10^6$ parameters, while most protein families contain 10^2 to 10^5 effective sequences (M_{eff}); this discrepancy grows quadratically as N increases.

To avoid overfitting to the limited training data, we applied the standard approach of adding a penalty term to the optimization problem that drives parameters to zero unless they explain a sufficient amount of the data. We follow Ekeberg *et al.*[180] and Kamisetty *et al.*[182] by employing l_2-regularization on the fields $\boldsymbol{h}$ and the pair couplings $\boldsymbol{J}$ with parameter type-specific regularization strengths $\lambda_{\boldsymbol{h}}$ and $\lambda_{\boldsymbol{J}}$, i.e.

$$\mathcal{R}(\boldsymbol{h}, \boldsymbol{J}) = \lambda_{\boldsymbol{h}} \sum_{i=1}^{N} \|\boldsymbol{h}_i\|_2^2 + \lambda_{\boldsymbol{J}} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \|\boldsymbol{J}_{ij}\|_2^2. \tag{2.9}$$

The optimization problem to be solved becomes a trade-off between maximizing the probability of the data (log of the pseudo-likelihood $\mathcal{L}(\boldsymbol{h}, \boldsymbol{J})$) and minimizing the complexity of the model (regularization term $\mathcal{R}(\boldsymbol{h}, \boldsymbol{J})$):

$$\arg\max_{\boldsymbol{h}, \boldsymbol{J}} \left(\log \mathcal{L}(\boldsymbol{h}, \boldsymbol{J}) - \mathcal{R}(\boldsymbol{h}, \boldsymbol{J}) \right). \tag{2.10}$$

In a Bayesian framework, this approach can be interpreted as maximum *a posteriori* inference under Gaussian priors on the parameters with a mean of zero and variances proportional to the inverse of $\lambda_{\boldsymbol{h}}$ and $\lambda_{\boldsymbol{J}}$[182].

Different values of $\lambda_{\boldsymbol{h}}$ and $\lambda_{\boldsymbol{J}}$ have been proposed by empiric optimization of contact prediction accuracy on sets of protein families[180,182]:

Method	$\lambda_{\boldsymbol{h}}$	$\lambda_{\boldsymbol{J}}$
Ekeberg *et al.*[180]	$0.01 \cdot M_{\text{eff}}$	$0.01 \cdot M_{\text{eff}}$
Kamisetty *et al.*[182]	0.01	$0.2 \cdot (N-1)$

Notably, the regularization approach by Kamisetty *et al.*[182] adjusts for the higher number of coupling parameters $\boldsymbol{J}$ compared to fields $\boldsymbol{h}$, and applies weaker regularization as the number of training samples growths (no dependence on the number of effective sequences M_{eff}). For structure predictions (membrane proteins and protein complexes), we used the approach by Ekeberg *et al.*[180], whereas for mutation effect predictions we applied the method by Kamisetty *et al.*[182]. Both approaches perform similarly regarding contact prediction accuracy, but initial experiments showed that the method by Kamisetty *et al.*[182] gives superior performance when predicting mutation effects.

We also explored the use of different types of regularization, including (i) group l_1-regularization which enforces sparsity on entire $\boldsymbol{J}_{ij}$ matrices but performs l_2-regularization on the elements within[181], as well as (ii) pair-specific weights $\lambda_{\boldsymbol{J}}(i, j)$ for l_2-regularization to adjust regularization strength based on site entropy. These initial explorations did

not result in improved prediction accuracy over standard l_2-regularization and are not reported in this thesis.

Parameters of the statistical model were estimated using a customized version of the code provided by Ekeberg *et al.*[180].

2.3. Phenotype prediction from coevolution model

The inferred probability model quantitatively describes evolutionary constraints on the composition of observed functional protein sequences. By interpreting the model and its parameters, we identified patterns on the genotype level and attempted to predict corresponding phenotypes on the molecular and organismal level. This includes the (i) calculation of summarized constraints betweens pairs of positions to obtain three-dimensional structure information and (ii) the calculation of mutation effects by relating the probabilities of sequences under the model.

2.3.1. Constraints between pairs of positions

For each pair of positions i and j, the coupling matrix $\boldsymbol{J}_{ij}$ describes the family-specific preferences for all possible amino acid configurations σ_i and σ_j. The larger the differences between entries of such a matrix are, the more one pair configuration is favored over the other, and the larger the epistatic constraint between the pair is. To quantify and compare the overall sequence-independent epistatic constraint between pairs of sites, the 21^2 numbers per $\boldsymbol{J}_{ij}$ matrix need to be summarized into a single measurement that is comparable between different protein families.

Calculation of evolutionary couplings

Previous work has addressed the problem of summarizing the coupling matrices using a mutual information-based measure called *direct information*, which is defined as the difference entropy between pair probabilities in a $\boldsymbol{J}_{ij}$-derived two-site probability model and the independent expectation from marginal single-site amino acid frequencies[8,173,174].

We focused on a matrix norm-based solution[173,175,180] that has been shown to give improved contact prediction accuracy compared to direct information[180]. In this approach, each coupling matrix $\boldsymbol{J}_{ij}$ is first centered around row and column means of zero by transformation into a *zero-sum gauge* using

$$\boldsymbol{J'}_{ij}(k,l) = \boldsymbol{J}_{ij}(k,l) - \boldsymbol{J}_{ij}(\cdot,l) - \boldsymbol{J}_{ij}(k,\cdot) + \boldsymbol{J}_{ij}(\cdot,\cdot) \tag{2.11}$$

where $\cdot$ means average across these entries, and then scored by calculating the Frobenius norm

$$FN(i,j) = ||\boldsymbol{J}_{ij}||_2 = \sqrt{\sum_k \sum_l \boldsymbol{J'}_{ij}(k,l)^2}. \tag{2.12}$$

which sums across all 21^2 amino acid combinations k, l. The transformation in Equation 2.11 minimizes the Frobenius norm[173] and, by shifting each coupling matrix to a zero mean, allows to interpret $FN(i,j)$ as a quantity proportional to the sample standard deviation of the matrix entries with a factor of $\sqrt{\frac{1}{21^2-1}}$.

We then applied the empirically derived *average product correction* (APC) to the FN matrix to remove background coupling between positions that arises due to confounding factors such as finite sampling and phylogenetic relationships between samples[171,175,180]. Assuming that, on average, each position should only be coupled to a small subset of all sites, the correction approximates the background coupling of both sites by the row and column averages of the matrix $(\cdot)$ and removes these from the raw score $FN(i,j)$, i.e.

$$EC(i,j) = FN(i,j) - \frac{FN(i,\cdot)\, FN(\cdot,j)}{FN(\cdot,\cdot)}. \tag{2.13}$$

This correction is identical to setting the largest eigenvalue of the FN matrix to zero and reconstituting the matrix from its eigenvectors (J. Söding, personal communication).

The end result of the calculation is a symmetric $N \times N$ matrix (N=length of protein) with entries estimating the strength of *evolutionary coupling* (EC) between all pairs of sites, where larger positive values correspond to high evolutionary coupling and values around zero correspond to no detectable evolutionary coupling.

Selection of significant evolutionary couplings

The evolutionary couplings calculation assigns an EC score to every pair of positions (i,j) in the protein. Yet, we do not expect all pairs of positions to be coupled, so we are left with the problem of identifying significant pair couplings in a comparable manner across different proteins. To select significant pair couplings in a scale-free way, we developed a strategy based on the following empiric observations:

1. While most pairs have an EC score around zero, there is a one-sided tail of positive scores containing a much lower fraction of pairs.

2. The higher the EC score of a pair in the tail is, the more likely it is to be proximal in 3D, while a large fraction of pairs in the background distribution is distant.

3. The background (non-tail) part of the distribution is approximately symmetric around a zero mean.

We therefore assumed that the more an EC score in the tail exceeds the background score distribution, the more likely it is to reflect "true" coevolution between the pair of positions. We used the minimal EC score between any pair to approximate the width of the symmetric background distribution and measured raw EC reliability as the ratio by which an EC score is above the background, i.e.

$$Q_{\text{raw}}(i,j) = \frac{EC(i,j)}{\min\limits_{i,j}(EC(i,j))} \tag{2.14}$$

Although Q_{raw} allows to make relative statements about EC scores within a protein, it is not directly comparable across different proteins as EC accuracy depends on the available amount of sequence information relative to the number of parameters of the statistical model. To correct for these factors, we applied an empirically derived normalization to the raw score, yielding the normalized quality score

$$Q(i,j) = \frac{Q_{\text{raw}}(i,j)}{1 + \left(\frac{M_{\text{eff}}}{N}\right)^{-\frac{1}{2}}} \tag{2.15}$$

where M_{eff} is the effective number of sequences after reweighting (Equation 2.8) and N is the number of sites in the statistical model. The correction strongly reduces the normalized reliability in the case of limited samples, but returns the raw reliability score in the limit of large amounts of training data.

For the special case of assessing protein complex interactions, we restricted Equation 2.14 to inter-monomer EC only and exclude any intra-monomer ECs from the calculation.

2.3.2. Three-dimensional structures of proteins and complexes

The evolutionary pressure to maintain functional proteins and protein interactions leads to the coevolution of amino acid residues in structural contact. Assuming that evolutionary couplings between positions reflect this pressure, structural phenotypes could be reconstructed from evolutionary sequence variation by constraining structure in 3D space using pairwise distance restraints (Figure 2.1)[4,8,9].

We applied this principle in two different ways: (i) the *de novo* 3D structure prediction of single protein molecules from extended amino acid polypeptides, based on further developments of our earlier work[4,8,9], and (ii) the prediction of protein complex interactions through docking assembly of predefined monomer protein structures.

Monomer structure folding

To obtain all-atom 3D models of proteins from evolutionary couplings, we further developed our previously described methods EVfold and EVfold-transmembrane[8,9].

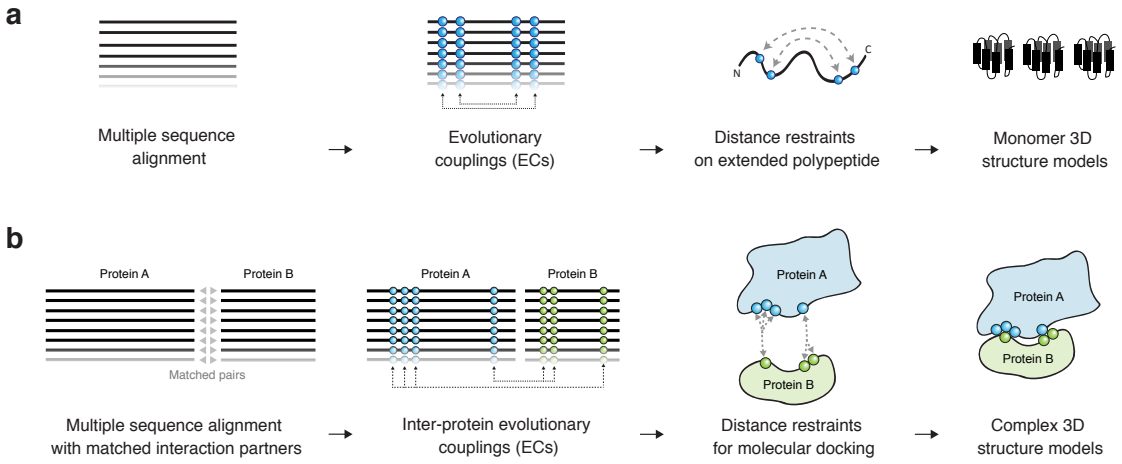

Figure 2.1.: Three-dimensional structures predicted from evolutionary couplings. (**a**) Starting from a sequence alignment of the protein family, evolutionary couplings between sites in the protein are calculated using the pairwise maximum entropy model. Under the assumption that coupled pairs are in physical contact, the distances of residues are restrained to reconstruct 3D structure models from an extended polypeptide. (**b**) For the prediction of complex structures, the sequences of interacting proteins need to be matched first for the calculation of inter-protein evolutionary couplings. Assuming proximity of the coupled residues between the two proteins to define distance restraints, the 3D structure of the complex can be reconstructed with molecular docking. Adapted from Hopf *et al.*[1,2]

Given a selected list of evolutionary couplings, the steps of the folding protocol are as follows:

1. *Filtering of ECs using sequence-based predictions:* Pair couplings not consistent with predicted secondary structure features because of geometric criteria (impossible contacts between residues in transmembrane segments, helices or beta strands) were excluded from structure prediction to increase the amount of ECs that most likely correspond to structural contacts and to exclude false positive couplings. Transmembrane segments were predicted with PolyPhobius[206] and compared to MEMSAT-SVM[207] and TOPCONS[208] predictions to assess reliability by consensus. Secondary structure was predicted using PSIPRED[209]. The precise filtering rules are described in full detail in our previous work[8,9].

2. *Generation of 3D models from ECs:* Folded all-atom 3D candidate models were generated by restraining the distances of EC residue pairs in a fully extended polypeptide. The embedding in 3D was performed using distance geometry and simulated annealing protocols of the Crystallography and NMR system (CNS)[210]. Additional distance and angle restraints were added based on local secondary structure including transmembrane helix segments, and long loop regions without EC coverage were removed from folding to avoid interference with constrained regions. For each set of evolutionary couplings, we sampled 20 candidate structures as the folding protocol can get stuck in local optima.

3. *Blind ranking of candidate models:* Benchmark experiments showed that the generated candidate models can have strongly varying quality and therefore need to be blindly assessed to select high-quality predictions[8,9]. We used the default EVfold-transmembrane quality assessment protocol that scores each model for its agreement with predicted secondary structure, predicted lipid exposure as well as agreement with additional evolutionary couplings[9]. To assess convergence to a common solution in 3D space and to exclude high-scoring outlier solutions, additional structure-based clustering was applied to filter the list of results.

The resulting ranked list of candidate 3D structures can be further validated using prior biological knowledge (e.g. crosslinking data, mutational studies) to assess the accuracy of structural models in the absence of experimental structures.

Protein complex docking

The coevolution analysis of protein interactions results in two sets of evolutionary couplings: intra-monomer ECs defining the structure of each of the subunits, and inter-monomer ECs determining the interaction between the subunits. Structural models of protein complexes could be derived from this information analogously to monomer structures by joint prediction of the monomer structures with additional distance restraints on the inter-molecular interactions. A simpler approach to the problem, which was used in this work, is to use pre-defined structures for the monomers (experimental or predicted) and assemble them into a 3D complex structure using protein-protein *docking* with EC-derived distance restraints.

To dock single monomers into complexes, we employed the HADDOCK[154] software and defined unambiguous distance restraints on the C_α atoms (5 ± 2 Å) for high-scoring inter-ECs with a normalized quality score ≥ 0.8 (Equation 2.15). The quality score threshold was derived based on an empiric trade-off between inter-EC precision and coverage (see Section 3.2.1 for details).

The HADDOCK docking protocol consists of three stages: (i) rigid-body energy minimization, (ii) semi-flexible refinement in torsion angle space, and (iii) model refinement in explicit water solvent. These stages go from a faster, coarse-grained search of 3D space to slower, fine-grained refinements allowing subtle conformational changes and amino acid side-chain rearrangements. We used default parameters for the protocol, but lowered the number of generated models during each stage to (i) 500, (ii) 100, and (iii) 100. We chose this reduction to demonstrate the targeted information added by ECs compared to sampling of a much larger search space by using energy functions alone. To test docking in the absence of EC information (negative control), we used the *ab initio* mode of HADDOCK[211] to calculate a larger number of models only with center of mass restraints enforcing contact between the subunits (for each stage: (i) 10000, (ii) 500, (iii) 500 models). The resulting set of candidate 3D models was scored and ranked using the default HADDOCK score[154], but excluding the

distance restraint energy term in the third stage (Edist3 = 0) to assess model quality independently of the ECs used to generate the models.

An important consideration for evaluating docking performance is the use of *unbound* monomer structures. If *bound* subunit structures are taken directly from the co-crystallized complex, the complementarity of conformations and side-chains may lead to an over-optimistic assessment of performance as the energy function can simply assemble the subunits in a more targeted search space. For a more realistic evaluation, we therefore took independently solved, unbound structures where available. In all other cases, we randomized the monomer side chains using SCWRL4[212] or Schrödinger Protein Preparation Wizard[213].

2.3.3. Mutation effects

The inferred probability model (Section 2.2), through site- and pair-specific amino acid constraints, describes how compatible sequences are with the functional requirements of the aligned protein family. The probability distribution could therefore in principle be used to quantitatively relate changes in genotype (mutations of amino acid sequence) to changes in phenotype, and ultimately, the fitness of the organism. Contrary to most previously published approaches for mutation effect prediction, the coevolution model allows to incorporate the context-dependence of mutations in the calculation using pairwise interactions between positions in the protein (Figure 2.2). While the focus of this work is on mutation effects in single proteins, we highlight that the calculations outlined in this section straightforwardly extend to protein interactions.

Probabilistic calculation of context-dependent mutation effects

To quantify mutation effects, we adopted a formalism used to predict changes in protein stability[139] that is based on the Boltzmann form of the distribution $P(\boldsymbol{\sigma}) = \frac{1}{Z}\exp\left(E(\sigma)\right)$ relating the energy E of a system configuration $\boldsymbol{\sigma}$ to its probability. For our pairwise model (Equation 2.1), the statistical energy of an amino acid sequence $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_N)$ is defined as

$$E(\boldsymbol{\sigma}) = \sum_{i=1}^{N} \boldsymbol{h}_i(\sigma_i) + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \boldsymbol{J}_{ij}(\sigma_i, \sigma_j) \tag{2.16}$$

and describes the favorability of the given system configuration by summing the specific single-site constraints $\boldsymbol{h}_i(\sigma_i)$ and pair constraints $\boldsymbol{J}_{ij}(\sigma_i, \sigma_j)$. We note that throughout this work we used a sign convention such that higher (rather than lower) statistical energies $E(\boldsymbol{\sigma})$ correspond to higher probabilities.

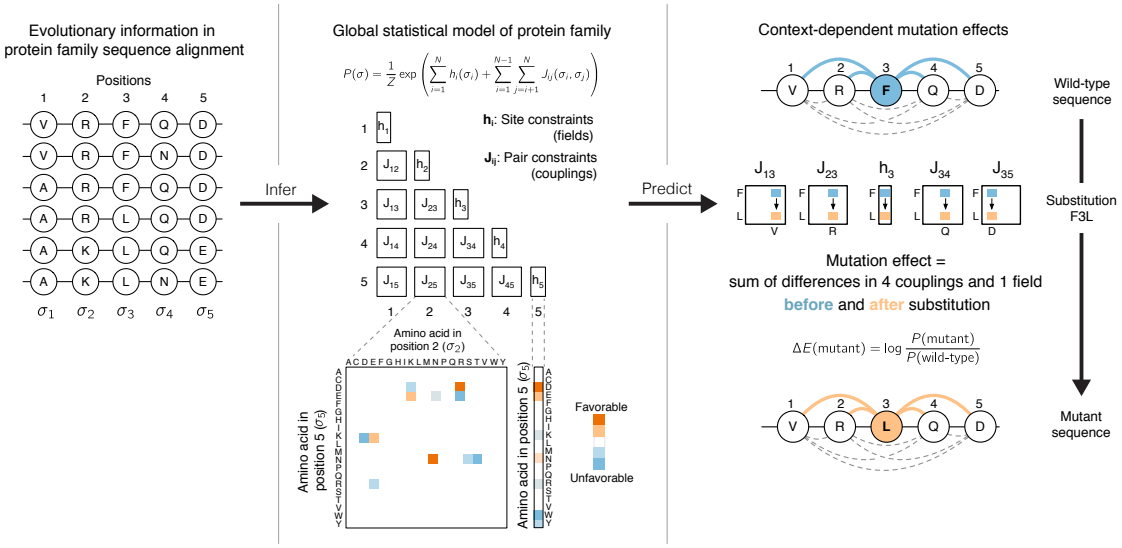

Figure 2.2.: Calculation of context-dependent mutation effects. *Left:* Evolutionary sequences record the functional constraints on the amino acid configurations in a specific protein family. *Middle:* By inferring a pairwise maximum entropy model from these sequences, we extract amino acid constraints on single sites (h_i) and pairs of positions (J_{ij}), which allows to compute the probability of any sequence to be a functional member of the protein family. *Right:* The probability distribution over sequences can be used to quantify the effects of mutations by relating the probabilities of the wild-type and mutant sequences under the model. Through the difference in pair couplings to all other positions, the method incorporates the sequence background in the calculation and in this way models epistatic dependencies. Adapted from Hopf *et al.*[3]

We calculated the effects of mutations as the statistical energy difference between the mutant and the original wild-type sequence,

$$\Delta E(\boldsymbol{\sigma}^{(\mathrm{mut})}, \boldsymbol{\sigma}^{(\mathrm{wt})}) = E(\boldsymbol{\sigma}^{(\mathrm{mut})}) - E(\boldsymbol{\sigma}^{(\mathrm{wt})}), \tag{2.17}$$

which is equivalent to the log-odds ratio of the probabilities of the two sequences (Figure 2.2):

$$\Delta E(\boldsymbol{\sigma}^{(\mathrm{mut})}, \boldsymbol{\sigma}^{(\mathrm{wt})}) = \log \frac{P(\boldsymbol{\sigma}^{(\mathrm{mut})})}{P(\boldsymbol{\sigma}^{(\mathrm{wt})})}. \tag{2.18}$$

The statistical energy difference quantifies how compatible the mutated amino acids are with (i) site-specific amino acid requirements through the differences of fields in the mutated sites, as well as (ii) the rest of the sequence through the differences in pair couplings to all other positions. The use of pair couplings allows to calculate context-dependent mutation effects rather than analyzing patterns of conservation in the mutated sites only.

If $\Delta E > 0$, the mutant is more likely under the model than the wild-type sequence (beneficial mutation); if $\Delta E < 0$, it is less likely (deleterious mutation). Mutants with $\Delta E = 0$ are equally likely (neutral mutation). Under the log-odds ratio formalism,

statistical energy differences can be intuitively interpreted as relative changes in probability. Between different proteins, varying scales of the predicted mutant effects may reflect different overall mutational susceptibilities/fitness costs for the organism. Here, to enable comparisons between different proteins, we discard this information and normalize predictions to a scale from -1 (deleterious mutation) to 0 (neutral mutation) using the transformation

$$\Delta E_c(\boldsymbol{\sigma}^{(\mathrm{mut})}, \boldsymbol{\sigma}^{(wt)}) = \frac{\Delta E(\boldsymbol{\sigma}^{(\mathrm{mut})}, \boldsymbol{\sigma}^{(wt)})}{|D|}. \tag{2.19}$$

$|D|$ is the mean statistical energy difference of the 5% most deleterious single mutants and is used as an approximation of the maximally deleterious single mutation effects that is more robust to outliers.

Due to the inclusion of the sequence context in the calculation, mutation effects computed using Equations 2.16, 2.17 and 2.19 are referred to as from the "epistatic model".

Calculation of mutation effects with independent site model

Mutation effects are frequently predicted based on single-site conservation without considering epistasis. To test if a context-dependent model captures mutation effects more accurately, for comparison we constructed a maximum entropy model that only contains first-order terms $\boldsymbol{h}_i$ characterizing positional amino acid constraints. Under this model, the probability of a protein sequence $\boldsymbol{\sigma}$ is

$$P(\boldsymbol{\sigma}) = \frac{1}{Z} \exp\left(\sum_{i=1}^{N} \boldsymbol{h}_i(\sigma_i)\right). \tag{2.20}$$

The parameters of the models were inferred from the sequence data using standard maximum likelihood estimation, which in this case is tractable due to the independence of sites. Analogously to the pair model, we applied l_2-regularization with strength $\lambda_h = 0.01$ during parameter learning to avoid overfitting (Section 2.2.2).

To calculate statistical energy differences between a mutant and wild-type sequence, the probabilities of both sequences according to Equation 2.20 were plugged into Equation 2.18 and rescaled with Equation 2.19. Here, the energy difference quantifies how compatible the mutated amino acids are with site-specific amino acid requirements without considering the rest of the sequence. Mutation effects computed based on Equation 2.20 are referred to as from the "independent model".

Quantification of context-dependence of mutation effects

Depending on the particular protein family, target sequence and position, epistasis may play a role in shaping the effects of mutations to a greater or lesser extent. We

used the prediction difference between the epistatic and the independent model to measure how context-dependent mutation effects are by subtracting the rescaled log-odds ratios, i.e.

$$\Delta\Delta E_{\text{ind}}^{\text{epi}}(\boldsymbol{\sigma}^{(\text{mut})}, \boldsymbol{\sigma}^{(\text{wt})}) = \Delta E_{\text{c}}^{\text{epi}}(\boldsymbol{\sigma}^{(\text{mut})}, \boldsymbol{\sigma}^{(\text{wt})}) - \Delta E_{\text{c}}^{\text{ind}}(\boldsymbol{\sigma}^{(\text{mut})}, \boldsymbol{\sigma}^{(\text{wt})}). \tag{2.21}$$

Mutants with $\Delta\Delta E_{\text{ind}}^{\text{epi}} < 0$ are predicted as less fit by the epistatic model than by the independent model, and vice versa. The higher the absolute value of $\Delta\Delta E_{\text{ind}}^{\text{epi}}$ is, the more the effect of a mutation depends on the particular sequence context into which it is introduced. When comparing the two models against experimental data, assessing those mutants that are predicted the most differently is of particular interest to gauge if epistatic models describe mutation effects more accurately than independent models.

2.4. Evaluation of phenotype predictions against experiments

We have described methods to predict protein phenotypes from evolutionary sequence covariation on the level of protein structures, protein complexes, and amino acid mutation effects. To assess the predictive performance of these methods, we evaluated the predicted phenotypes against the available experimental phenotype measurements. In the following section, we describe how representative data sets were chosen and how predictive accuracy was measured between predictions and experiments.

2.4.1. Dataset selection

For each of our phenotype prediction applications, we searched for available experimental data to obtain comprehensive evaluation sets, subject to the availability of sufficient amounts of evolutionary sequence information for the respective target proteins.

Membrane protein structures

Performance of the updated version of our method for membrane protein structure prediction was evaluated on a dataset described previously[9]. Briefly, we selected 25 target proteins from 23 non-redundant α-helical membrane protein families with five or more transmembrane segments, sufficient sequence information, and at least one solved 3D crystal structure in the PDB database[78]. For a separate set of 17 (out of a total of 18) proteins without any available experimental structure for the entire protein family, we provided confident *de novo* prediction models in 2012[9]. Crystal structures have been published for four of these proteins with confident predictions since[214–217], and we evaluated predictive accuracy for these on the original[9] and updated predictions.

Protein complex structures

To test the accuracy of inter-monomer ECs and docked 3D models, we derived a protein complex structure data set based on a collection of experimentally verified binary protein-protein interactions in *E. coli*[218]. This collection was assembled from yeast two-hybrid experiments, literature curation and three-dimensional structures of complexes in the PDB database[78], and subsequently expanded by us to contain additional missing candidates[1]. The full list of binary interactions, which still contained complexes without 3D information, was then filtered by the following criteria:

1. Availability of a high-quality 3D crystal structure containing both monomers in co-crystallized form, either directly for the interacting proteins or between two homologous proteins as identified by individual *jackhmmer*[192] searches against the PDB for each monomer. Structures were required to cover at least 30 residues per monomer, and to have a resolution smaller than 5Å.

2. Proximity of the interacting pair on the *E. coli* genome, since genomic distance was used as a proxy to match up putatively interacting sequence pairs across species (Section 2.1.3). Proximity was measured by the number of genes in between the two interacting monomers as defined by an ordered list of genes on the *E. coli* genome obtained from UniProt[202]. Complexes with a gene distance greater than 20 were excluded from the evaluation set.

3. Interaction of monomers belonging to two distinct protein families. Protein complexes, where both subunits belonged to the same Pfam[219] protein family ('pseudo-homomultimers'), were excluded from the evaluation set, since alignment construction and disentangling inter- and intra-ECs requires a different approach due to the common evolutionary origin of both subunits (e.g., false signal for inter-ECs based on intra-subunit couplings).

This filtering procedure yielded a set of 93 complex structures that could in principle be applied for evaluation purposes. When constructing alignments, aligned regions were restricted based on the crystal structure if the statistical inference problem became too large regarding computational resource usage or the amount of available sequence information. The other remaining 229 interactions without 3D structure of the complex but satisfying requirements 2 and 3 were used as potential *de novo* prediction candidates.

Mutation effects

In contrast to the well-defined evaluation targets of three-dimensional structure prediction, data on the phenotypic effects of mutations are often heterogeneous: (i) Mutant effects are assessed based on a variety of assays and interrelated target phenotypes, including organism growth, protein stability, enzymatic activity and many others; and (ii) effects are typically tested only on a small subset of mutated positions

and amino acid substitutions based on prior information such as patterns of sequence conservation[18,60,167,220].

To circumvent this biased coverage of mutational space, we focused our evaluation on using quantitative high-throughput mutagenesis scans characterizing mutation effects on different phenotypic readouts[60]. We performed a comprehensive search of the literature for such scans that cover entire proteins or protein domains, and excluded any experiment for which the target protein did not have sufficient available evolutionary sequence information (redundancy-reduced number of sequences $M_{\text{eff}} < 10N$, N=length of protein/domain). In total, 13 mutation scan data sets covering 11 unique proteins were identified and used for evaluation (Table A.1)[41,44,51,53,54,59,60,62,64,66,68–71].

As the resolution of high-throughput experiments is limited due to threshold and saturation effects[59], predictions were also tested against selected low-throughput measurements of protein stability and enzyme activity from biochemical studies focused on protein sequence coevolution[107,167,220].

2.4.2. Measures of predictive accuracy

Given experimentally determined phenotypes, the quality of the corresponding predictions was assessed by scoring the agreement between both using established measures of predictive accuracy.

3D structures of proteins and complexes

The quality of protein structure phenotype predictions was quantified in two different spaces: (i) The agreement of predicted residue-residue contacts with the contacts observed in the experimental structure, and (ii) the deviation between EC-predicted and experimental atomic coordinates in 3D.

The predictivity of evolutionary couplings for residue-residue contacts was assessed by their *precision*, i.e. what fraction of evolutionary couplings corresponds to residue pairs that are close (true positives) rather than distant (false positives) in the 3D structure of the protein:

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}} \tag{2.22}$$

Precision calculations were restricted to the highest-scoring EC pairs, since only a small fraction of all possible pairs is assumed to capture significant coevolution (Section 2.3.1). Typically, we used all pairs with EC quality score above a certain threshold or the top L (length of protein/domain) couplings as previously described[8,9], and excluded pairs (i, j) that were close in the primary amino acid sequence ($|i - j| \leq 5$). Depending on the particular application, two residues were classified as close (in contact) if the Euclidean distance between any pair of atoms was no more than 5 Å (monomer protein structures) or 8 Å (inter-monomer contacts in protein complexes). We note that

due to the particular nature of the structure prediction problem, the identification of relatively few reliable and informative contacts is more important than to obtain the full set of close pairs or to predict pairs that are distant[8]; hence the focus on precision as the main evaluation metric.

While residue-residue contacts can be evaluated in a binary classification setting, the assessment of structure predictions against experimental structures requires to compare the coordinates of two three-dimensional objects. A standard way of comparison is the root mean square deviation (RMSD), which scores the distances between equivalent atoms in an optimal superposition of prediction and model that minimizes the RMSD[221], i.e.

$$\text{RMSD} = \min \left\{ \sqrt{\frac{1}{N} \sum_{i=1}^{N} d(i)^2} \right\}. \tag{2.23}$$

For two structures with a total of N aligned atom pairs, $d(i)$ measures the Euclidean distance between the coordinates of the i-th pair. Structure predictions of single proteins were evaluated based on superpositions of all modeled C_α atoms. For protein complexes, where the focus is on accurate modeling of the interaction interface region, we used all backbone atoms of interface residues (interface RMSD)[222,223]. The set of interface residues was defined to contain all residues with any atom closer than 6 Å to any atom of the interaction partner.

A second evaluation metric we used for monomer proteins is the Template Modeling (TM) score, which addresses the tendency of the RMSD to increase with the size of the aligned molecules[224]. Similar to the RMSD, it is based on an optimal superposition that maximizes the score for a target protein of length L and N aligned atom pairs:

$$\text{TM-score} = \max \left\{ \frac{1}{L} \sum_{i=1}^{N} \frac{1}{1 + \left(\frac{d(i)}{d_0(L)} \right)^2} \right\} \tag{2.24}$$

The function $d_0(L) = 1.24\sqrt[3]{L - 15} - 1.8$ normalizes the atom pair distances $d(i)$ such that the composite score is independent of the length of the protein L. On the scale of the TM score from 0 to 1, values of approx. 0.5 indicate that both proteins are likely to share the same fold, while larger scores correspond to increasingly good agreement between the compared structures[225].

Structure comparisons were performed using MaxCluster[ii], PyMOL[226], ProFit[iii] and TM-align[227].

[ii] www.sbg.bio.ic.ac.uk/~maxcluster/
[iii] www.bioinf.org.uk/software/profit/

Mutation effects

To assess if quantitative mutation effects computed from sequence variation (Section 2.3.3) correspond to quantitative experimental measurements, we used correlation measures to quantify the dependence between the two variables. Since initial visual inspection of the relationship between predictions and experiments showed the presence of a variety of linear and non-linear dependencies and marginal distributions of different shapes, we simultaneously applied several established measures to obtain a robust characterization of the results:

1. Pearson product-moment correlation coefficient r to measure the linear dependence between prediction and experiment. Pearson's r is directly related to the coefficient of determination r^2 of the corresponding linear regression problem with intercept term, which in our setting can be interpreted as the percentage of variance in the experiments explained by evolutionary sequence variation. The Pearson correlation may give misleading results if the data have a non-linear relationship or outliers are present[228].

2. Spearman rank correlation coefficient ρ to capture monotonic, non-linear dependencies between prediction and experiment, while being robust to the presence of outliers. Spearman's ρ is equivalent to calculating Pearson's r on the respective ranks of the data points in the two distributions[228].

3. Matthews correlation coefficient (MCC)[229] to test prediction performance in a binary classification setting, i.e. if experimentally deleterious/neutral mutations are correctly predicted as deleterious/neutral or not. The binary analysis was motivated by the observation that many of the analyzed experimental datasets show a bimodal effect distribution and most mutations are either very deleterious or neutral. For a 2×2 contingency table containing the counts of true positive (TP), false positive (FP), false negative (FN) and true negative (TN) prediction instances, the MCC is defined as

$$\mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}. \tag{2.25}$$

   To calculate the MCC, experimental mutations were classified as deleterious or neutral by fitting a two-component Gaussian mixture model to the effect distribution in logarithm space and then assigning each mutation to the class with the higher posterior probability under the model. As predicted effect distributions tended to be unimodal, there was no obvious cutoff to separate neutral and deleterious mutants. The MCC was therefore evaluated at a range of different thresholds for the predictions.

All three correlation measures range from $r = 1$ for a perfect correlation to $r = 0$ in the absence of correlation to $r = -1$ for a perfect anti-correlation.

2.5. Data analysis and visualization

Mutation effect calculations, data analysis, statistical calculations and visualization of results were performed using the scientific Python stack[230–235] and IPython notebooks[236]. Protein structures were visualized using PyMOL[226]. Mappings between UniProt sequences[202] and structures in the PDB[78] were obtained from the SIFTS project database[237].

3. Results and Discussion

In the following chapter, we will outline the results of predicting transmembrane protein structures, protein-protein interactions and quantitative context-dependent mutation effects from evolutionary sequence covariation. For each of these target phenotypes, we describe the experimental data used for the evaluation of the respective prediction method and how well benchmarking predictions agree with these experiments. We then apply the methods to obtain *de novo* predictions of phenotypes for proteins lacking experimental information.

3.1. Transmembrane protein structures

The structures of α-helical transmembrane proteins are a molecular phenotype of particular interest since membrane-integral proteins are responsible for intercellular communication and substrate uptake, and therefore major drug targets[238]. In previous work, we have shown that transmembrane protein structures can be accurately predicted from sequences and provided *de novo* predictions for proteins for which there was no three-dimensional structure available[9].

Here, we evaluate the updated version of our original prediction method (Chapter 2) on the same benchmark dataset and assess if the updates lead to more accurate predictions. For some of the unsolved proteins predicted in Hopf *et al.*[9], experimental structures have been published in the meantime, providing an excellent opportunity for a completely blinded evaluation of the original predictions in the spirit of the CASP (Critical Assessment of protein Structure Prediction) competition[239].

We then proceed to predict the three-dimensional structures of insect olfactory receptors (ORs), a unique family of proteins responsible for the translation of environmental chemical signals into neuronal activity[240]. The OR family, which has distant homologs in non-insect animals and plants, appears to be of ancient evolutionary origin and does not share any detectable similarity to existing protein families with solved structure[241]. We characterize the validity of our models using independent, orthogonal experimental data and targeted experiments based on our predictions.[i]

3.1.1. Evaluation of updated prediction method

We first tested the performance of our updated prediction method in comparison to the original method[9]. The original approach was modified in several steps, including the generation of sequence alignments using jackhmmer (Section 2.1.1), calculation

[i] This section is based on the publication by Hopf *et al.*[2] and unpublished results.

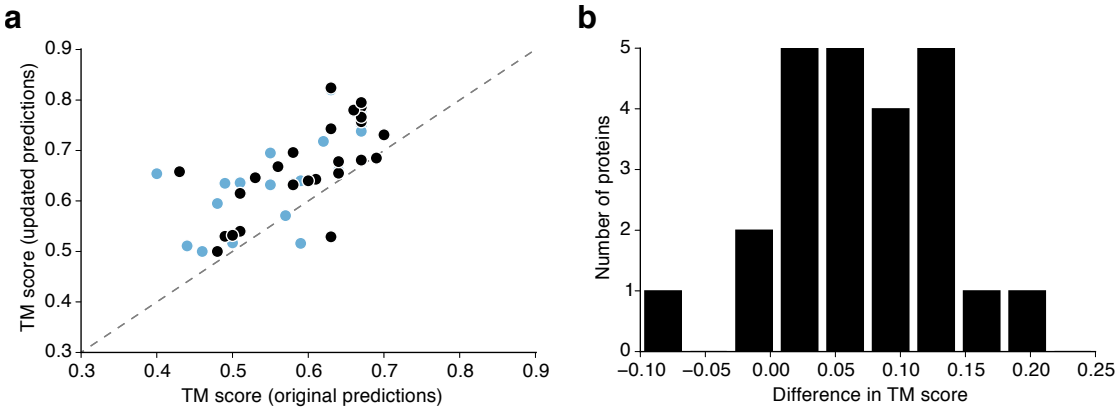

Figure 3.1.: Updated membrane protein method gives more accurate predictions. (**a**) Structure predictions using the updated method are equally or more accurate than those from the original method for both the best (black) and top-ranked models (blue, best out of 10 top-ranked models) with the exception of one protein. (**b**) Close to two thirds of the tested proteins show a substantial improvement over the old predictions (difference in top-ranked TM score $\geq$ 0.05).

of evolutionary couplings using the pseudo-likelihood maximization approximation (Section 2.2.2) and default topology predictions based on PolyPhobius (Section 2.3.2). The amount of available sequence information has more than doubled since the initial publication of our original method to over 40 million sequences[202], which could improve predictive accuracy given its strong dependence on the number of sequences observed previously[9,182].

Updated method improves model accuracy on original benchmarking set

We computed updated predictions for the 25 proteins described in the original benchmark set[9] and identified the models that were the most accurate compared to the experimental structure (*best model*) and had the highest blind quality assessment score (*top-ranked model*, accuracy reported for best out of 10 top-ranked models; Section 2.4.2)[ii]. For all but one protein from the data set (ferrous-iron efflux pump FieF), the updated method returned equally or more accurate predictions (Figure 3.1). On average, TM scores improved by 0.07 for the best models and by 0.08 for the top-ranked models, giving average TM scores of 0.67 (best) and 0.66 (top-ranked), respectively. None of the updated top-ranked models had a TM score smaller than 0.5, indicating that the correct fold was blindly identified for all proteins; the most accurate generated model for any protein had a TM score as high as 0.82 (*E. coli* NADH-quinone oxidoreductase subunit N, NuoN). In total, 8 out of the 25 proteins now had top-ranked models with TM scores of at least 0.7, and 18 out of 25 were above 0.6.

[ii]The updated benchmark and *de novo* predictions described in this section were computed by Li Yang Smith in Debora Marks' lab at Harvard Medical School co-advised by Thomas A. Hopf

Table 3.1.: Evaluation of experimentally solved *de novo* structure predictions

			Top-ranked TM score[a]	
Predicted protein	UniProt ID	3D Structure	Original	Updated
Adiponectin receptor 1	ADR1_HUMAN	3wxvA[214]	0.69	0.79
NADH-ubiquinone oxidoreductase chain 1	NU1M_HUMAN	4he8C[215] (39%)[b]	0.50	0.73[c]
Solute carrier family 22 member 4	S22A4_HUMAN	4pypA[216] (23%)[b]	0.57	0.80[c]
Translocator protein	TSPOA_HUMAN	2mgyA[217] (81%)[b]	0.52	0.61[c]

[a]TM score for best out of 10 top-ranked structures [b]% sequence identity of target protein to solved homologous structure
[c]Updated prediction calculated for solved protein rather than original target

In summary, our results show that the updated method improves the accuracy of three-dimensional structure predictions of membrane proteins from sequences and gives reliable results for all tested proteins. These developments are in line with original extrapolations that ongoing method development and more available sequences will lead to more accurate models and more accessible families in the near future[4,182].

Predictions for unsolved proteins have correct fold

As part of the original publication in 2012[9] and additional web supplemental data, we predicted three-dimensional models for 18 proteins where no structural information was available for the entire protein family (models available on evfold.org) and obtained reliable predictions in 17 cases. Since then, experimental structures of the target protein or one of its homologs have been solved for 4 of the 17 proteins. Similar to the CASP competition, this allows to assess the performance of the original method on fully blinded predictions in a retrospective evaluation[239].

For all four solved proteins, we compared our original predictions to the solved crystal structures; if a homolog from the family was solved, we used SwissModel comparative modeling[242] and evaluated our predictions against the homology model. In all four cases, the best out of the 10 top-ranked models had a TM score of at least 0.5 (Table 3.1), suggesting that our predictions captured the correct overall fold (TM score ≥ 0.5) but with mistakes in the structural details. In particular, we correctly predicted that the first subunit of the NADH-ubiquinone oxidoreductase is structurally similar to the other membrane subunits of the same complex despite a lack of detectable sequence similarity[9]. The experimental structure of the full respiratory complex 1 later confirmed our prediction[215]. Similarly, our structural model of the human adiponectin receptor 1 indicated that this protein shares the same fold as G protein-coupled receptors and bacterial rhodopsins but with inverted membrane topology[9]. Comparison against the solved experimental structure confirms this observation and shows that our prediction had substantial accuracy (TM score: 0.69).

To test how well our updated method would have predicted the unsolved proteins, we also computed new predictions for all four proteins (now directly for the solved structures rather than the original proteins). The TM scores of the updated top-ranked models increase considerably by up to 0.23 (Table 3.1). In part, these differences may be attributed not only to improvements in modeling performance, but also to the comparison against the native structure rather than a homolog.

The successful blind prediction of proteins that had not been solved at the time, as well as the increase in performance with the updated method suggest that the approach could be confidently applied to predict the structures of unsolved proteins of interest.

3.1.2. De novo model of insect olfactory receptors

Insect olfactory receptors are a large family of α-helical membrane proteins involved in the molecular recognition of odors. Despite great interest in elucidating the precise signaling mechanism of these ion channels, the molecular details of how olfactory receptors detect specific ligands and transmit this signal across the membrane remain elusive[243]; one reason for this is the absence of three-dimensional structure information for the protein family. As a contribution towards a more detailed understanding of olfactory receptor function, we predicted their structure and analyzed the resulting models in the context of existing biological knowledge and a targeted experimental validation.

Insect ORs have a unique seven-helix fold

To predict the three-dimensional structure of olfactory receptors, we followed the approach outlined in Chapter 2 and built a custom sequence databases of 5907 OR sequences that were already deposited in public databases, or obtained from newly sequenced insect genomes to maximize the amount of available sequence information. We chose two of the experimentally most well-characterized members of the OR family, the co-receptor ORCO and the ligand-specific receptor OR85b from *Drosophila melanogaster*, as target proteins for our predictions and built multiple sequence alignments for both proteins (Section 2.1.1). For each of the alignments, we calculated evolutionary couplings (Section 2.3.1), generated three-dimensional models from EC-derived distance restraints, and blindly selected a top-ranked candidate model that had high quality scores and clustered with other high-ranking solutions (Section 2.3.2).

For both proteins, the high-scoring evolutionary couplings show agreeing parallel and anti-parallel interaction patterns between membrane-integral helices typical for α-helical transmembrane proteins when visualized as a contact map (Figure 3.2a)[9]. While similar results are expected for both proteins since they belong to the same family and their sequences can be aligned over most of their length, this convergence to the same typical patterns nevertheless hints at the robustness of the calculated evo-

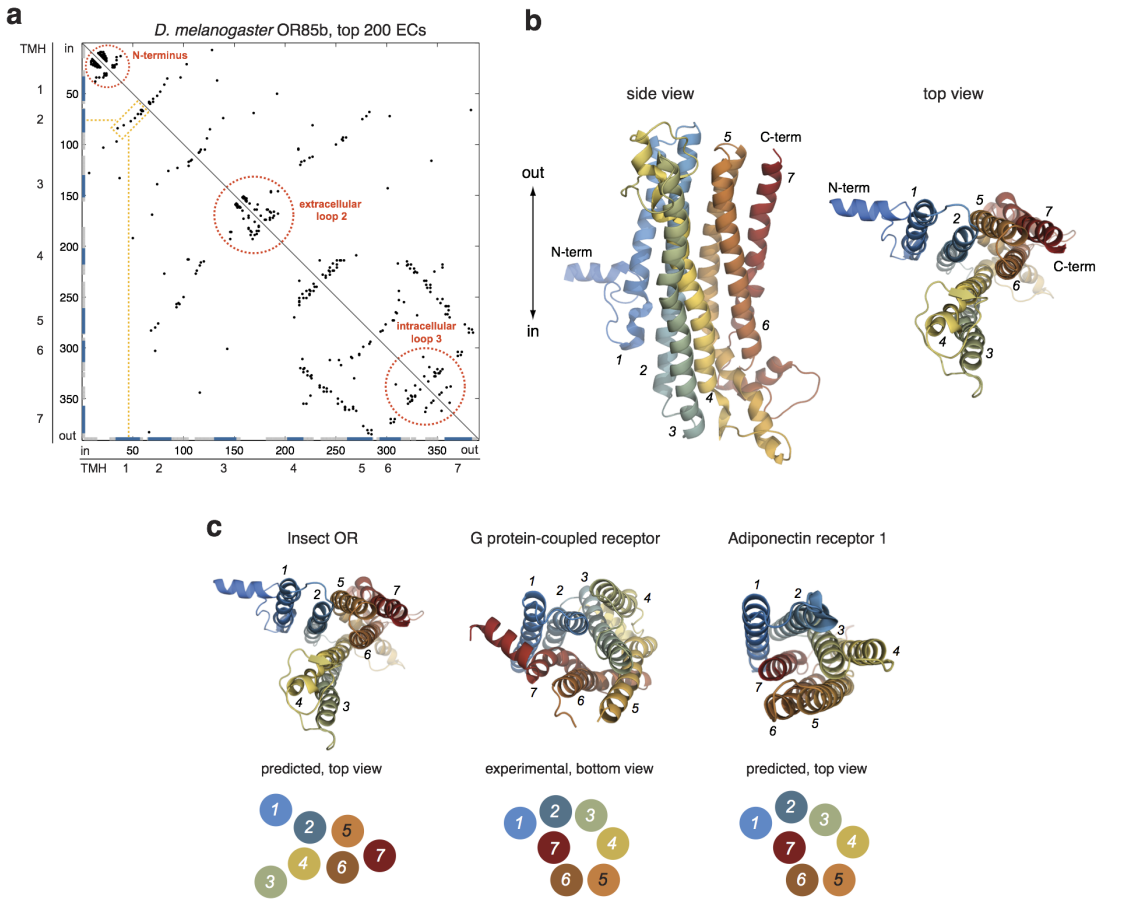

Figure 3.2.: Structural model of insect olfactory receptors from evolutionary couplings. (**a**) Contact map of the highest-scoring evolutionary couplings between positions (x- and y-axis) in the olfactory receptor OR85b. Parallel and anti-parallel stretches of couplings typical for α-helical transmembrane proteins can be observed between helices spanning the membrane (blue segments, TMH 1-7) and outside the membrane (grey segments), e.g. between TMH 1 and 2 (orange rectangle). The N-terminal tail region, extracellular loop 2 and intracellular loop 3 display high local densities of evolutionary couplings (red circles). (**b**) Top-ranked 3D model of OR85b viewed from within the membrane (side view, left panel) and from the extracellular face of the membrane (top view, right panel; blue to red coloring from N- to C-terminus). (**c**) The helical packing arrangement of insect ORs (OR85b, top-ranked model) is distinct from G-protein coupled receptors (β2-adrenergic receptor, PDB structure 2rh1[244]) and the adiponectin receptors (top-ranked model from Hopf *et al.*[9], confirmed by PDB structure 3wxv[214]), although all three have seven transmembrane helix segments (top panels: view extra- or intracellular face of membrane; bottom panels: simplified representation of helical packing; blue to red coloring from N- to C-terminus). Adapted from Hopf *et al.*[2]

lutionary couplings. The selected top-ranked structures, although differing in their structural details (C_α-RMSD=5.9 over 303 residues, TM-score=0.54, calculated using TM-align[227]), share the same overall three-dimensional packing of the seven transmembrane helices of OR proteins (Figure 3.2b), revealing a coarse-grained model of the olfactory receptor fold for the first time.

Based on the observation that ORs have seven transmembrane helices, an open question in the field of olfactory receptor research has been if this class of receptors belongs to the family of G protein-coupled receptors (GPCRs)[245,246]. GPCRs are an abundant class of receptors with seven transmembrane helices in eukaryotes and responsible for olfactory reception in non-insect species[240]. A comparison of packing arrangement shows however that the topologies of ORs and GPCRs differ substantially (Figure 3.2c), suggesting that these are in fact unrelated protein families and folds; the topology is also different from that of the adiponectin receptor which has a fold similar to GPCRs but with inverted orientation inside the membrane[9]. Additional structure-based similarity searches of the top-ranked models against the PDB[78] using DALI[247] did not lead to the identification of any other membrane proteins sharing the same fold, highlighting that ORs have a novel and currently unique three-dimensional structure.

Strongly coupled and known functional residues cluster in 3D model

Given the calculated evolutionary couplings and the resulting candidate models, we are left with the questions about the validity of the model and what residues contribute to the signaling function of insect olfactory receptors. We therefore used our previously described strategy of identifying residues with above-average coupling strength as a proxy to predict potentially functionally important residues[9] and validated the results against experimental data. Briefly, this strategy is based on the assumption that residues with strong evolutionary couplings to other positions are under particular selective constraint, which is possibly not visible from single-site conservation alone, and thus hypothesized to be functionally important. For each position, we (i) calculated the weighted degree in the network with positions as nodes and evolutionary coupling scores between pairs of positions as edge weights, and (ii) normalized this score by the average edge weight of any position; (iii) positions with a normalized score > 1 have above-average coupling strength (referred to as *strongly coupled positions*). The mapping of strongly coupled positions onto the three-dimensional structure models (Figure 3.3a) reveals that they cluster in three distinct parts of the molecules, suggesting functional importance of these regions: (i) the N-terminal tail, (ii) the second extracellular loop (EL2) connecting the third and fourth transmembrane helices (TMH3/TMH4); and (iii) the third intracellular loop (IL3) and the subsequent seventh transmembrane helix (TMH7).

Several independent experimental mutational studies have tested the functional contributions of selected residues to receptor function for different ORs[248–256]. We compiled a list of such functional studies and mapped residues that contribute to OR

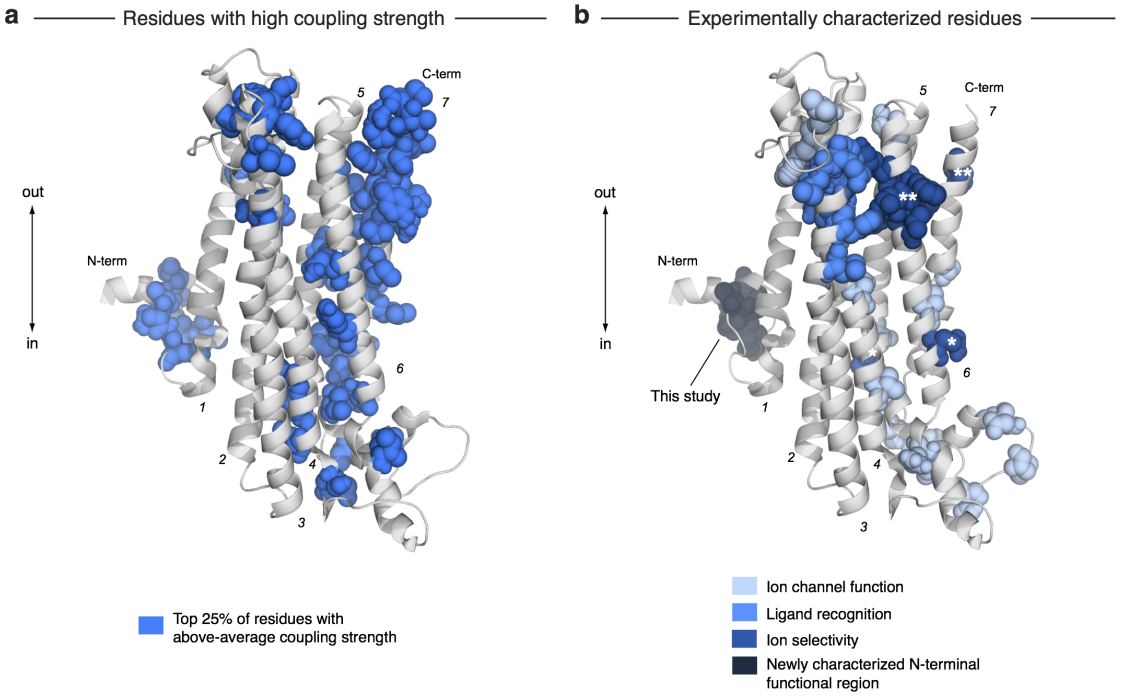

Figure 3.3.: Strongly coupled positions in olfactory receptors cluster in regions with experimentally tested contributions to function. (**a**) Positions in OR85b with above-average evolutionary coupling strength cluster in the N-terminal region, extracellular loop 2 and intracellular loop 3/transmembrane helix 7 (top 25% mapped on top-ranked OR85b model as blue spheres; side view from inside the membrane). (**b**) Experimentally characterized functional positions in members of the OR family on the predicted structure of OR85b (top-ranked model, mapping based on sequence alignment of characterized proteins). Residues with defined influence on ion selectivity are all located in transmembrane helices 5, 6 and 7 and proximal in 3D (single and double asterisks). Adapted from Hopf *et al.*[2]

function onto OR85b and ORCO using our sequence alignments of the OR family (Figure 3.3b). Although the exact overlap between these and strongly coupled positions is limited to only two positions (N143 and F380), the experimentally tested positions cluster in distinct regions of the model that largely correspond to strongly coupled regions (Figure 3.3a,b). Interestingly, mutations affecting ligand recognition (TMHs 2-4, EL2) are spatially close on the extracellular side, suggesting they constitute part of the ligand-recognition site. Similarly, mutations on TMHs 5, 6 and 7 affecting ion selectivity are in spatial proximity in the model (Figure 3.3b, single and double asterisks); mutations affecting general ion channel function are mostly located on one face of the model in TMHs 5, 6, 7 and IL3, suggesting this part of the molecule forms the pore of the ion channel.

Experiments verify predicted functional role of N-terminal region

Since there was only anecdotal experimental evidence for the functional importance of the N-terminal region[248], our collaborators experimentally tested the role of the strongly coupled positions in this part of the molecule by generating a higher-order ORCO mutant substituting the strongly coupled residues A23, M24, F30, M31, H32 and N33 to alanine, as well as an ORCO deletion construct lacking residues 23 to 33. In experimental assays measuring odor-evoked current responses of ORCO to three different agonists (pentyl acetate, 2-heptanone, VUAA1), the two mutants had diminished or abolished receptor function compared to wild-type ORCO (for full experimental details, see Hopf *et al.*[2]). Although the precise functional role of the N-terminal region has yet to be uncovered, our computational predictions guided the identification of its essential contribution to signaling. Taken together, these results indicate a model of the spatial organization of functional regions in ORs and support the plausibility of our three-dimensional model.

3.1.3. Discussion

Based on the computation of evolutionary couplings from sequence covariation, we have predicted the previously uncharacterized three-dimensional structure of insect olfactory receptors. Besides other early examples[257–259] that followed our *de novo* predictions for several unsolved protein families[9], this work constitutes one of the first targeted applications of evolutionary couplings to study the biology of a particular protein of interest and highlights the power of approaches combining computational predictions and wet-lab experiments.

While our validation suggests that the coarse-grained models are reasonably accurate and can contribute to the further elucidation of structure-function relationships in this protein family, there are however several issues which need to be considered when interpreting the predictions. First, the models could be inaccurate due to prediction errors accumulating throughout the different stages of the method. The identified evolutionary couplings may contain false signals due to insufficient sequence information or other peculiarities of the protein family, the same applies for the machine-learning based predictions of secondary structure and membrane topology; wrong or missing pair restraints may then bias 3D reconstruction towards the wrong answer. Future work will therefore need to establish additional independent quality assessment criteria to determine how reliable results are. The performance of our updated method on the benchmark set as well as the successful predictions for proteins that have been solved in the meantime however give basic confidence in the applicability of the approach to ORs.

Second, ORs have been shown to form hetero-complexes of the co-receptor ORCO and a ligand-specific receptor such as OR85b in an unknown stoichiometry[260]. Although there is no detectable evolutionary relationship on the sequence level, struc-

tures of other ion channels have shown intertwined arrangements of the subunits in the complex[261] . Since ORCO and the other ORs are part of the same family, evolutionary couplings may occur either due to intra-subunit or inter-subunit residue-residue couplings; disambiguation between the two is a challenging problem for future research. During the folding of monomer structures as performed here, the presence of inter-molecular couplings can potentially give incorrect results. If ORs perform major conformational changes, the averaging of couplings caused by multiple conformations could also lead to an averaging of the different conformations in the predicted model[9].

Finally, the current approach averages the covariation signal across most of the OR sequences known today so there are enough samples for learning the statistical model. Due to the divergent nature of the OR family[241], this may lead to a loss of specificity in modeling the precise details of subfamilies, e.g. potential structural divergence between ORCO co-receptors and the ligand-specific receptors.

We anticipate that with the ongoing sequencing efforts, potentially including tens of thousands of different insect species, the repertoire of known OR sequences will continue to increase. Together with the continued development of coevolution-based methods, future prediction attempts could result in higher-resolution models to elucidate the structural and functional details of this and many other unsolved protein families.

3.2. Protein-protein interactions

One level up from individual protein structures, protein-protein interactions are molecular phenotypes relevant to most cellular processes. Although high-throughput screens have identified large sets of binary protein interactions, their molecular details such as the mode of binding, binding specificity, and conformational changes remain largely unknown[150,218,262]. To address this gap from sequence data, we developed a method to infer coevolving pairs of positions between two different proteins (*heterocomplexes*). Using experimentally solved complex structures as a benchmark, we show that evolutionarily coupled residues often correspond to inter-protein structural contacts, which contain enough information to reconstitute the overall three-dimensional structure of the complex from its subunits. We then demonstrate that the method can be applied to predict unsolved protein complexes, yielding biologically plausible results.[iii]

3.2.1. Benchmark of method on solved complexes

We tested our method for the identification of coevolving residues between interacting proteins using a dataset of 93 different non-redundant bacterial complexes with known structures where both subunits are located proximally on the genome (Section 2.4.1). For all complexes, we generated sequence alignments for both monomer proteins and matched putatively interacting pairs of homologs in the alignments using the genome distance-based strategy described in Section 2.1.3. Discarding paired alignments with insufficient sequence information ($M_{\text{eff}}/N < 0.3$, where M_{eff} the is effective number of sequences and N is the total number of positions in the matched alignment), we were left with a set of 76 complexes for which evolutionary coupling scores could be calculated. The calculation simultaneously returns EC scores for all pairs of positions between both proteins (*inter-protein* ECs) as well as for pairs of positions within the proteins (*intra-protein* ECs). To test if inter-protein ECs capture the coevolution of contacting residues, we checked the distances of high-scoring pairs in the available crystal structures of the complexes.

Coevolving pairs between proteins are close in structure

An initial evaluation of the benchmark set showed that top-ranked inter-protein ECs correspond to structural contacts in the complex crystal structures for many of the tested examples, but the fraction of correctly identified contacts strongly varies from case to case when picking a fixed number of contacts (Supplementary File 1 of appended publication[1]). To blindly select reliable EC pairs while discarding non-significant couplings, we developed a quality score that allows to set an expectation for the desired precision of the chosen contacts based on the overall distribution of EC scores (described in detail in Section 2.3.1, Figure 3.4a).

[iii] This section is based on the publication by Hopf *et al.*[1]

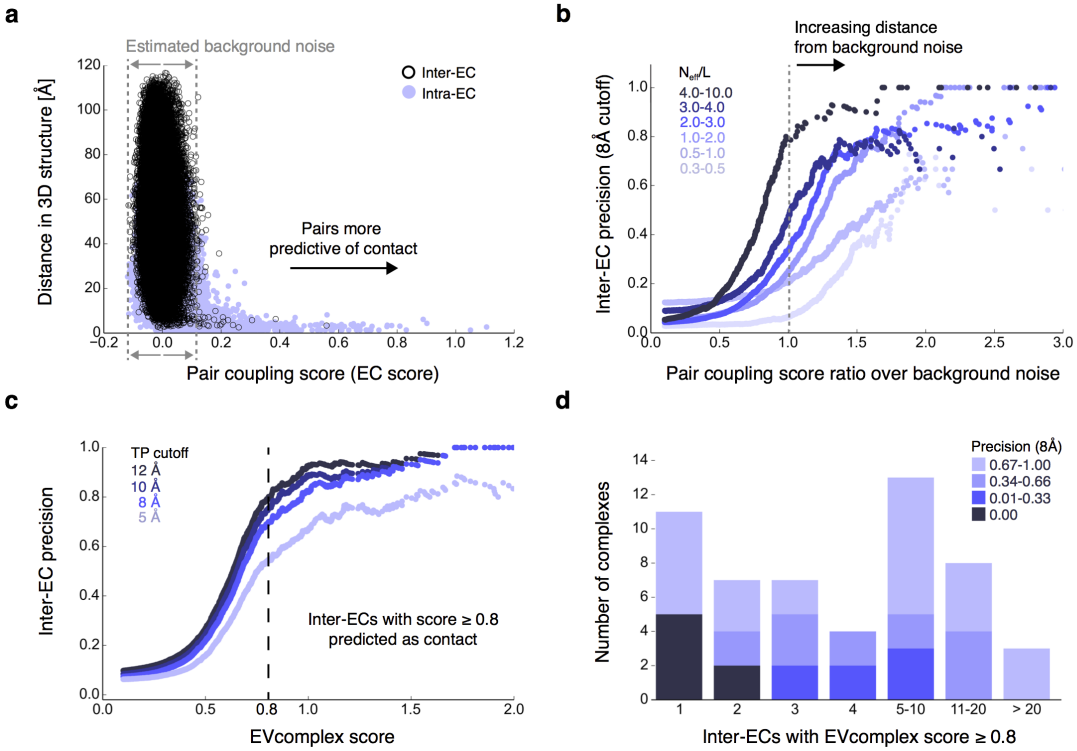

Figure 3.4.: Coevolving pairs in protein complexes correspond to structural contacts. (**a**) Residue pairs within the monomers (intra-ECs, blue spheres) and between the monomers (inter-ECs, black spheres) with high coupling scores are mostly proximal in the 3D structure of the complex (here: ABC transporter MetNI, PDB 3tui[263]). The largest fraction of EC pairs is distant in 3D with coupling scores distributed approximately symmetrically around 0 (background noise). (**b**) The more distant an inter-EC score is from the background noise (Equation 2.14), the more likely the corresponding pair is to be predictive of structural proximity in 3D. This relationship of normalized EC score to precision is dependent on the amount of available sequence information (curves in different shades of blue). Estimates are more accurate for complexes with higher amounts of sequence information, requiring smaller distances from the background noise to obtain the same level of precision (plot limited to range 0–3 to focus on phase transitions of curves). (**c**) Normalization of this score for the amount of available sequence information (Equation 2.15) allows to estimate the average expected EC precision for a given quality score threshold. In this work, we selected inter-ECs with score $\geq$ 0.8 as predicted contacts between proteins (plot limited to range 0–2). (**d**) Precision of inter-ECs as predictor of interacting residues for all complexes with at least one significant inter-EC (quality score $\geq$ 0.8, true positive contact defined at 8 Å minimum atom distance cutoff). Adapted from Hopf *et al.*[1]

The score measures (i) how much of an outlier an inter-protein EC pair score is compared to the background distribution of non-significant inter-protein couplings, (ii) normalized for the overall expected reliability of ECs based on the number of available non-redundant sequence data per position (M_{eff}/N). Without the normalization (step ii), the precision of selected ECs at any given reliability score threshold would be strongly dependent on the the amount of sequence data (Figure 3.4b). For example, a score of 1.0 or larger would correspond on average to approximately 80%

correctly identified contacts if M_{eff}/N is in the range from 4.0-10.0, but only 20% correctly identified contacts if M_{eff}/N ranges from 0.5 to 1.0. The empirically derived normalization step (Equation 2.15) corrects for this discrepancy and gives a quality score that is independent of the amount of sequence information for the particular complex (Figure 3.4c). Using the observed shape of the curve relating the quality score and the corresponding precision of inter-protein ECs for different definitions of structural contacts (in the range from 5 to 12 Å), we chose a threshold of 0.8 to select significant inter-protein ECs; we note that a trade-off between the number of identified contacts and their precision can be obtained by dialing through this curve.

Selecting all inter-protein ECs above the threshold of 0.8, on average 69% of these pairs are closer than 8 Å in the corresponding three-dimensional complex structures. There is however substantial spread in the number and accuracy of the selected EC pairs (Figure 3.4d; Figure 3.5a,b). For 53 of the 76 complexes, we chose at least one inter-protein EC. The majority of these 53 cases has more than one selected pair, in 24 cases, there are 5 significant inter-protein ECs or more. Generally, the more pairs pass the quality score threshold, the higher the overall precision of these pairs is (Figure 3.4d). A particularly striking example is the bacterial two-component signaling complex with 78 selected pairs of which 72% are correct, highlighting why this complex with abundant copies in bacterial genomes ($M_{\text{eff}}/N > 95$) has served as the primary application case for computational studies of inter-protein coevolution[93,173,177]. Remarkably, in some examples with detailed experimental studies available, the identified high-scoring inter-protein ECs correspond to known functional residues. For example, the top-ranked evolutionary couplings in the ATP-binding cassette (ABC) transporter MetI-MetN are part of a residue network coupling ATP hydrolysis in the ATP-binding domain to substrate transport in the membrane-integral domain (Figure 3.5a)[263,264]. This suggests that evolutionary coupled pairs are indicative not only of structural contacts, but also of critical functional constraints to maintain the function of the protein interaction.

However, for 23 of the 76 benchmark complexes no EC pair has a score above the chosen threshold; possible reasons why no significant coevolution was detected in these cases are discussed in Section 3.2.3.

Complexes can be accurately reconstructed with contact-based docking

Motivated by the observation that the coevolution analysis of protein complexes gives accurate inter-protein contacts, we then asked if this information could be used to reconstruct the three-dimensional structure of the complex from its subunits. We chose a representative set of 15 complexes with 5 or more inter-protein ECs above the quality score threshold, and assembled the monomer proteins into a complex by protein-protein docking with HADDOCK and distance restraints enforcing spatial proximity of the EC residue pairs (Section 2.3.2). To test how well docking works in the absence

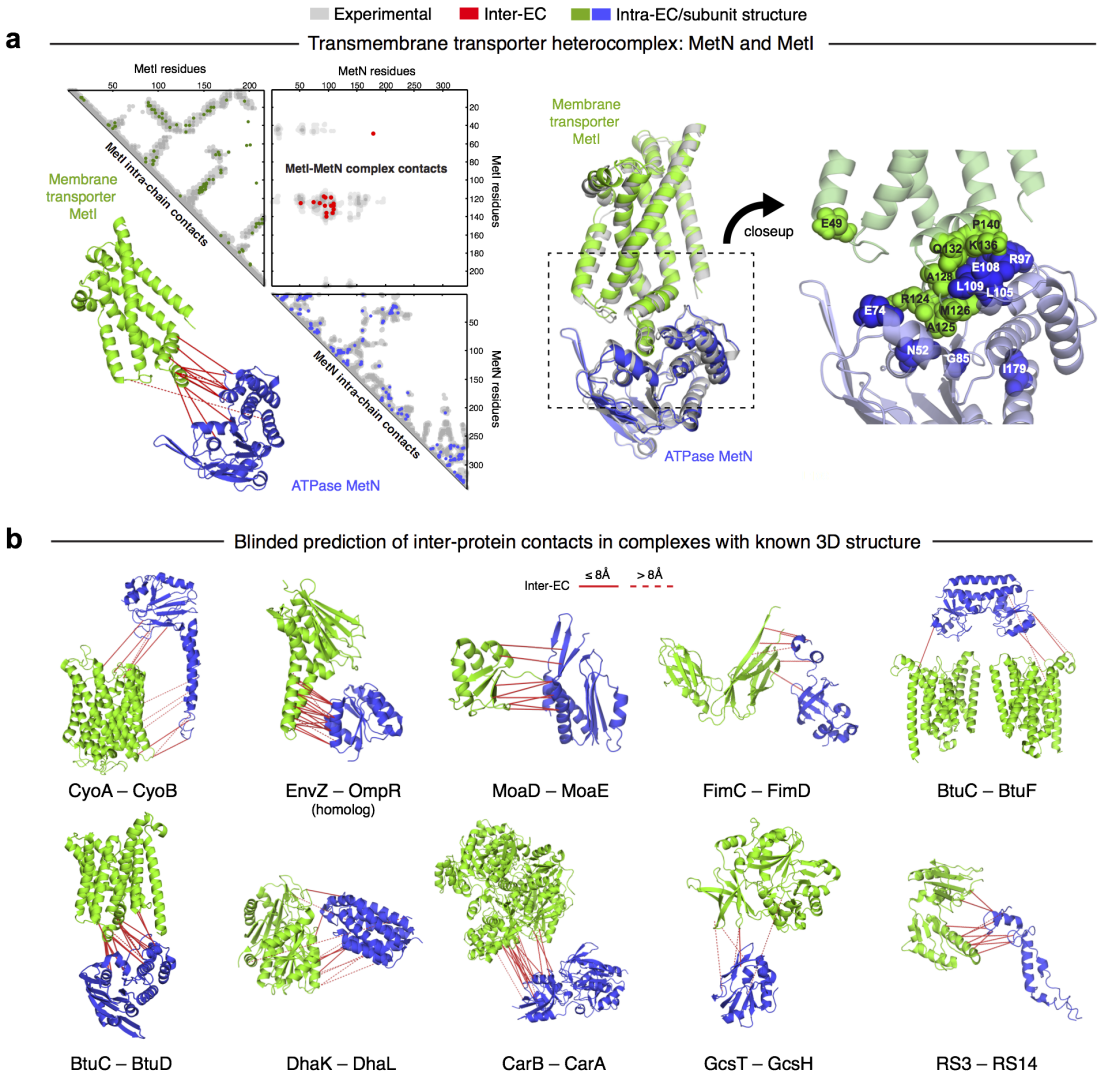

Figure 3.5.: Complex 3D structures predicted from evolutionary couplings. (**a**) *Left:* A contact map representation of significant inter-ECs (red dots, upper right quadrant) and the corresponding intra-ECs (green/blue dots, triangles) shows that predicted contacts largely correspond to proximal pairs in the 3D structure of the MetIN ABC transporter complex (PDB structure 3tui[263]; dark, medium and light gray dots for distance cutoffs of 5, 8, and 12 Å, respectively), defining the structural interaction between both subunits (red lines connecting green and blue cartoons; bottom left). *Middle:* Docking based on significant inter-ECs (top-ranked model, green and blue cartoon) accurately reconstitutes the complex as observed in the experimental structure (grey cartoon, PDB entry 3tui[263], 1.5 Å interface-RMSD). *Right:* Close-up of the complex interface region with residues coupled by significant inter-ECs (green and blue spheres). (**b**) Gallery of complexes with known experimental structures and at least 5 significant inter-ECs that were tested using docking (score $\geq$ 0.8; monomers as green and blue cartoons; true positive contacts as solid, false positives as dashed red lines). Adapted from Hopf *et al.*[1]

of EC information, we additionally generated negative controls without any restraints other than enforcing the subunits to assemble with center of mass restraints.

The docking process using inter-protein ECs consistently generated accurate models of the complexes when compared to their experimental structures. For all 15 complexes, the most accurate out of the 100 resulting candidate models has an interface RMSD within 5.5 Å to the experiment; in 8 of the examples, at least 80 of the 100 candidates are within 4.0 Å. The energy function reliably and blindly selects good candidate models, which is a strong requirement for useful *de novo* predictions: no selected model is worse than 7.2 Å, and in 10 of 15 cases the selected model is within 4.0 Å. In 7 blindly selected cases, we obtained interface RMSDs of 2.0 Å or lower; not surprisingly, these tend to be complexes with more than 10 identified inter-protein ECs and relatively high EC precision ($\geq$ 0.5 at 8.0 Å distance cutoff).

The docking results using EC-based distance restraints are in pronounced contrast to the negative controls with center of mass restraints. Despite exploring the space of possible solutions more for the control (500 vs. 100 generated models), only a very small fraction of candidates (for all complexes $<$1.2%) is within 4.0 Å of the experimental structure. The scoring function is however not able to reliably select these solutions blindly and the top-ranked model has an interface RMSD of 15.0 Å or more for 12 of the 15 tested complexes.

The enrichment of good models ($\leq$ 4.0 Å interface RMSD) for the EC-based docking experiments and the successful blind identification of such structures suggests that inter-protein ECs add substantial information to the prediction of the three-dimensional structures of protein complexes.

Evolutionary couplings predict protein interactions

So far, we tested if evolutionary couplings can detect structural contacts in known protein interactions. Can the approach also be used to predict *if* two proteins interact, in addition to *how* they interact?

We chose *E. coli* ATP synthase, an essential protein complex responsible for the generation of ATP molecules, as test case for this question (Figure 3.6a). ATP synthase consists of 8 different subunits that are located in the cytoplasm (α, β, γ, δ, ϵ) or in the membrane (a, b, c). Crystal structures and extensive cross-linking studies have revealed the details of most interactions between subunits in this complex, with only the three-dimensional structures of the interactions between the a, b and c subunits remaining elusive[265]. To test if there is a coevolutionary signal to discriminate which of the 8 subunits interact, we computed evolutionary couplings for all possible 28 pairwise interactions and predicted protein pairs to be in contact if there is at least one significant inter-protein EC at our default quality score threshold of 0.8 (see above and Section 2.3.1). At this threshold, 24 out of the 28 interactions are correctly predicted to be interacting or not; only 4 pairs are wrongly classified as non-interacting (β-ϵ, β-b, c-γ, c-ϵ) even though there is some experimental evidence for an interaction. By

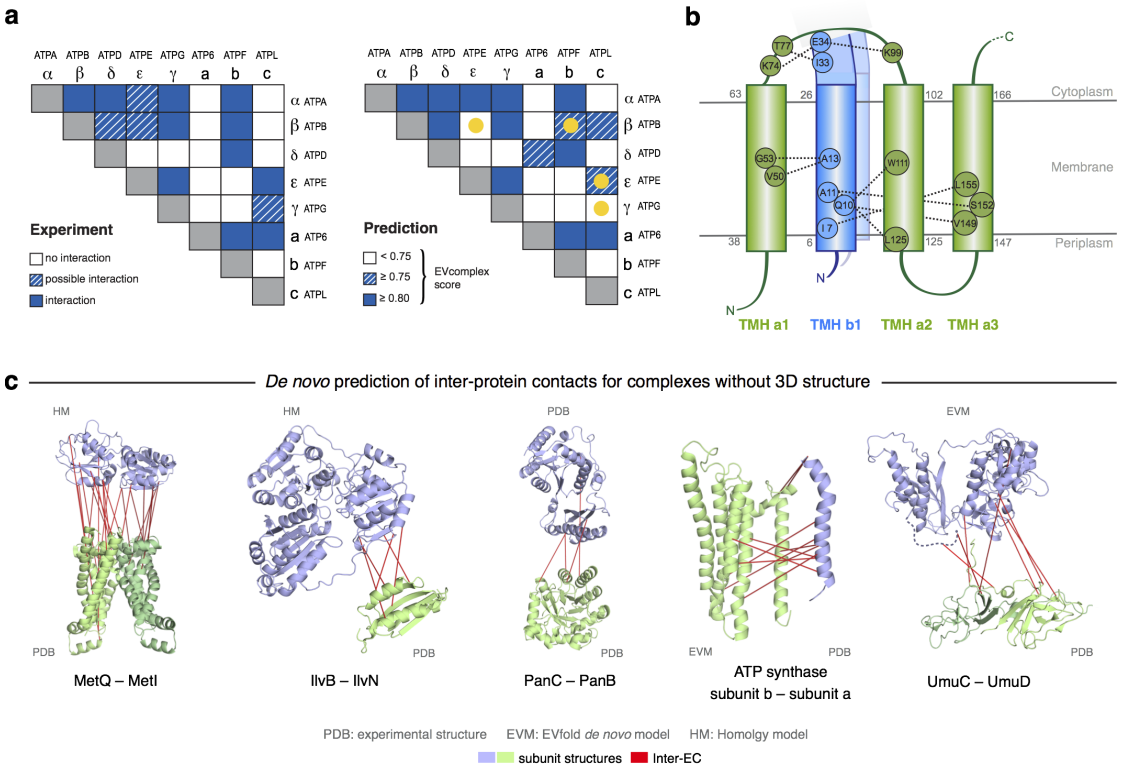

Figure 3.6.: De novo predictions of unsolved protein complexes. (**a**) Experimentally known (left) and predicted (right) interactions between subunits of *E. coli* ATP synthase. Four interactions have experimental evidence, but are not predicted using the default score threshold of 0.8 (false negatives, yellow dots). (**b**) *De novo* prediction of residue interactions between subunits *a* and *b* of ATP synthase (all significant inter-ECs with score $\geq$0.8). (**c**) Significant *de novo* predictions of residue interactions for complexes of unknown experimental structure (red lines: inter-ECs with quality score $\geq$0.8). Monomer structures (green and blue cartoons) were obtained from the PDB where available, or predicted using comparative modeling or evolutionary couplings otherwise; inter-ECs were distributed arbitrarily on the monomers for interaction partners that homomultimerize. Adapted from Hopf *et al.*[1]

lowering the quality score threshold to 0.75, two of the false negative predictions are classified correctly (β-*b*, *c*-ϵ) at the expense of introducing two false positives (β-*c*, δ-*a*).

The accuracy of the predictions for ATP synthase indicates that this coevolutionary approach could be used on a bigger scale to detect protein-protein interactions on a genome-wide scale and to identify how large protein complexes assemble.

3.2.2. De novo prediction of unsolved complexes

The dataset of 3449 protein-protein interactions used to derive the benchmark set of 93 complexes also contained interactions for which no structural information was available[218]. To provide *de novo* predictions for the subset of cases where our method is applicable, we built sequence alignments for all complexes that are close on the genome

(229 complexes, monomers no more than 20 genes from each other) and selected the 82 examples with enough sequences ($M_{\text{eff}}/N \leq 0.3$) and no structure of interacting homologs (Section 2.4.1). Of these, 32 complexes had at least one significant inter-protein EC with quality score ≥ 0.8; 10 complexes had 5 or more couplings.

Evolutionary couplings give plausible *de novo* models

Our predictions could be a valuable resource for further studies of these unsolved complexes. To verify if the results are biologically plausible, we analyzed some of the predicted ECs in the context of prior biological knowledge (Figure 3.6b,c). For the analysis of structural plausibility for complexes with unsolved monomer proteins, we predicted these structures using SwissModel comparative modeling[242] where a solved homolog was available, and created *de novo* models using our coevolution-based structure prediction protocol otherwise[8,9].

The interaction with the highest number of significant inter-ECs (24 pairs) is between the membrane domain of the D-methionine transporter MetI and its periplasmic binding protein MetQ that delivers the substrate to the transporter[266]. Consistent with the known cellular localization of the complex, all of the top 15 inter-protein ECs are between residues located on the two-lobe face of MetQ that gives access to the ligand-binding site, and residues on the exposed periplasmic face of the MetI homodimer. Similar clustering patterns were observed for several of the other predicted complexes, including IlvB-IlvN, PanC-PanB, UmuC-UmuD and ATP synthase subunits *a* and *b* (Figure 3.6c). During the publication of these results, the structure of the bacterial toxin/antitoxin complex DinJ-YafQ (19 significant inter-protein ECs) was released, providing an excellent opportunity to verify our blind prediction against experimental data (PDB entry 4q2u[267]). When comparing the 19 significant couplings to the crystal structure, we found that 17 of the residue pairs are no more than 8 Å from each other in the experimentally observed complex tetramer arrangement. These findings highlight that our method can provide accurate interaction patterns for unsolved protein complexes of interest if the monomers are close on the genome and have enough sequences.

Evolutionary couplings provide insight into unsolved ATP synthase interactions

The ATP synthase complex in *E. coli* has been characterized in detail with several crystal structures[265]. However, both the monomer structure of the membrane-integral *a*-subunit as well as the complex structure of its interaction with subunits *b* and *c* remain unknown except for cross-linking studies and coarse-grained structural models derived from these results[265,268,269]. Since the ubiquitous conservation of ATP synthase throughout all kingdoms of life leads to the availability of abundant sequence information[265] and solved interacting pairs in our benchmark set could be predicted

successfully, these missing pieces presents a suitable opportunity to apply our approach.

We first predicted the monomer structure of the *a*-subunit using our protocol for α-helical membrane proteins (Section 2.3.2). The resulting model, a four-helix bundle of transmembrane helices 2 to 5, is consistent with experimental information about residue proximities[269] and earlier computational models derived from this data (PDB entry 1c17[268]). Interestingly, there is only weak coupling of transmembrane helix 1 to the rest of the protein, suggesting this helix is not packed against the four-helix bundle; this is in agreement with experimental studies that failed to detect cross-links between these two regions[270]. Structural information for the N-terminal, membrane-integral part of the *b*-subunit that presumably interacts with the *a*-subunit is also limited and consists of an NMR structure of the single molecule (PDB entry 1b9u[271]) and cross-linking-based models of its experimentally verified homo-dimerization.

We then analyzed the 10 significant inter-protein ECs between subunits *a* and *b* identified by coevolution analysis of this protein interaction (Figure 3.6b). All 10 pairs are between the membrane-integral part of subunit *b* (residues 1-34) and membrane helices 1, 2, 3 and 5 of subunit *a* as well as the cytosolic loop connecting helices 1 and 2. All of these interactions are supported by experimental cross-links of the coupled residues or their sequence neighborhood, indicating that the coupled residues are in fact close in 3D (Supplementary File 6 of appended publication[1]). The construction of an explicit three-dimensional model of the complex was however hindered by the overall geometry of the interaction, as we were not able to identify an arrangement of a single *a* subunit with a tightly packed *b* homo-dimer model that simultaneously satisfies the couplings to helices 2, 3 and 5 (Figure 3.6b). A model based on experimental cross-linking only described interactions between subunit *b* and helices 2 and 5 of subunit *a*[272]. Similarly, the couplings of helix 1 of subunit *a* to subunit *b* could hint at an intertwined arrangement of subunits *a* and *b*. Future work will need to address these questions in more detail to obtain a detailed structural model of this elusive protein-protein interaction.

3.2.3. Discussion

In this work, we have shown that interacting residue pairs in protein complexes can be predicted from sequence covariation alone and that this information is sufficient to reconstruct the three-dimensional structures of the complexes from their monomers with high accuracy, allowing the *de novo* prediction of unsolved examples. Together with independent parallel work by Ovchinnikov *et al.*[273] that reports very similar results, to our knowledge the work presented here constitutes one of the first systematic applications of global coevolution models to the prediction of protein interactions.

Matching of interacting protein pairs is main limitation of method

A major limitation of coevolution based approaches for protein complex prediction, as is evident from our results, is the requirement to obtain a sufficiently large sequence alignment of evolutionarily related proteins interacting in the same or a similar way (here called *interaction homologs*). Our results show a clear trend that the protein complexes with a larger number of effective sequences per residue tend to give more significant EC pairs which have higher precisions when evaluated against solved structures (Section 3.2.1).

The generation of such alignments is non-trivial since we need to detect coordinated exchanges between pairs of sequences that actually interact. However, we usually do not know (i) how conserved the protein interaction is across orthologs in different species and across paralogs of the two monomers within each species, and (ii) how to identify the interacting pairs if there are multiple possible pair combinations, which is further complicated by the possibility of non-specific/promiscuous (i.e. there is no 1:1 matching of interacting proteins) or cross-species interactions (e.g. host-virus). We focused on a heuristic strategy (Section 2.1.3) to partly address these two questions by assuming that two homologs interact if their genomic distance does not exceed a certain limit (conservation of the interaction on an operon) and they are mutually closest to each other on the genome (disambiguation of multiple pairs). If there are no paralogs of either interaction partner, this approach trivially selects the only possible combination.

Using this strategy, which is only applicable to bacteria because of the operon organization of their genome, we did not predict any significant inter-protein ECs for 23 of the 76 benchmark complexes with enough sequences despite the prior knowledge that these proteins interact. This observation suggests that the above requirements for sequence selection and matching have been violated, resulting in a loss or decrease of coevolutionary signals *between* the proteins while intra-protein ECs were still accurately captured. With several thousands of bacterial genomes sequenced, the current approach works well for obligate protein complexes that are present throughout all bacteria, or where multiple copies exist that can be correctly disambiguated by genomic proximity (in our dataset, less than 10% of approximately 3500 known interactions in *E. coli*). Complexes that are only found in a limited subset of bacterial species, in eukaryotes, or are non-obligate, may not be amenable to our approach until many more genomes are sequenced – thus alleviating the need to match paralogs – and the method is further refined. More elaborate approaches to sequence pair matching could include phylogenetic strategies[164], improved ortholog and paralog identification, or solutions that iteratively include and exclude sequence pairs starting from a high-confidence seed alignment.

Prediction of homomultimers remains an open challenge

This work explicitly focused on interactions between two different proteins that also belong to distinct families (heterocomplexes). Interactions may also occur between two members of the same family, or by homomultimerization of the same protein. The identification of homomultimeric structures from sequence coevolution is a fundamentally different problem to the one presented here: For heterocomplexes, the main challenge is to identify and match interacting sequence pairs; the distinction between inter- and intra-protein ECs is trivial because couplings can be unambiguously assigned to either class based on which positions they are between. For homocomplexes and their self-interaction, there is no need to match partners (one has to find however the proteins that maintain the same pattern of homooligomerization). The main challenge for homocomplexes is to disentangle which evolutionary couplings are due to intra-molecular coevolution of the monomer structure, and which are due to inter-molecular coevolution between the assembling monomers[9].

Initial work has addressed this problem by using experimental monomer protein structures to filter EC pairs that are in spatial proximity in the monomer or have low surface accessibility as intra-protein ECs, and keep the other high-scoring ECs as inter-protein ECs, allowing the accurate 3D reconstruction of the homooligomer structure[274]. This approach can provide useful information about homocomplexes, but its applicability is limited to cases where the monomer structure is known but not the oligomer. The development of approaches that can blindly distinguish between both types of ECs in the absence of structural information is a task for future research; possible solutions could include (i) more elaborate statistical models that may discriminate based on subtle differences in the patterns of the coevolutionary signal or (ii) specialized 3D reconstruction, e.g. by folding monomers and the complex at the same time and iterative assignment of ECs to either class based on observed violations in the 3D model or ambiguous restraint definitions, as applied in the NMR field[275,276].

Future research challenges

Our results demonstrate that detailed information about protein-protein interactions can be inferred from sequence coevolution. Owing to the technological limitations outlined above, the method currently is only applicable to a subset of all complexes which we anticipate could be partially solved in future work. A method that can be used on a larger scale would enable a genome-wide study of protein-protein interactions and their structural details, giving orthogonal information to existing experimental approaches.

Even with the complexes accessible today, coevolutionary analysis opens a window into the sequence determinants of the emergence and molecular specificity of protein interactions[91,93,277,278]. A quantitative understanding of these genotype-phenotype re-

lationships could eventually help to obtain a detailed view of signaling networks , their malfunction in disease, and allow the targeted design of protein complex formation[279].

3.3. Phenotypic effects of mutations

The prediction of the phenotypic effects of mutations is of key relevance to genetics, genomics and clinical applications (Section 1.1.1). Current methods are however largely limited to compute binary effects while ignoring the potential context-dependence of genetic variants. We developed a probabilistic method to compute quantitative effects of mutations to proteins from evolutionary sequence covariation, while incorporating epistatic interactions with other loci within the protein. To gauge if predicted effects from the model correspond to experimentally measured phenotypes, we evaluate the agreement between computed effects and deep mutational scanning experiments which systematically explore the relationship of genotype and phenotype changes. We demonstrate that the incorporation of epistatic interactions in the probabilistic model improves the agreement with experimentally tested phenotypes, particularly in functionally important sites defining ligand and interaction specificities.[iv]

3.3.1. Evaluation of predicted mutation effects against experiments

The developed model for mutation effect predictions relates the probabilities of the mutant and wild-type sequence under a pairwise maximum entropy model of the protein family. The calculation of these probabilities incorporates all possible interactions between pairs of sites, and therefore should be able to capture potential dependencies of amino acid substitutions on the overall sequence background up to second order (Section 2.3.3). To test if computed quantitative phenotype changes (*statistical energy*) upon mutation correspond to experimentally determined phenotype changes, we searched the literature for deep mutational scanning experiments of entire proteins or protein domains and selected those where the target protein was member of a sufficiently large protein family (Section 2.4.1)[41,44,51,53,54,59,60,62,64,66,68–71], as well as low-throughput measurements of protein stability and enzymatic activity[167,220]. We inferred probability models for each target protein, computed quantitative effects for all experimentally tested mutations and then assessed the agreement between prediction and experiment.

Coevolution model predicts experimental phenotype changes

From our literature search, we obtained a data set of 15 mutagenesis studies for 13 unique proteins from bacteria, eukaryotes and viruses (Table A.1). Using a wide array of assays, these experiments measured the consequences of protein sequence mutations on different *in vitro* and *in vivo* phenotypes including growth under environmental pressure, peptide binding or protein stability.

[iv] The work in this section has been performed in collaboration with Debora Marks, Chris Sander, Michael Springer, Frank Poelwijk and John Ingraham. A preprint has been published in Hopf *et al.*[3]

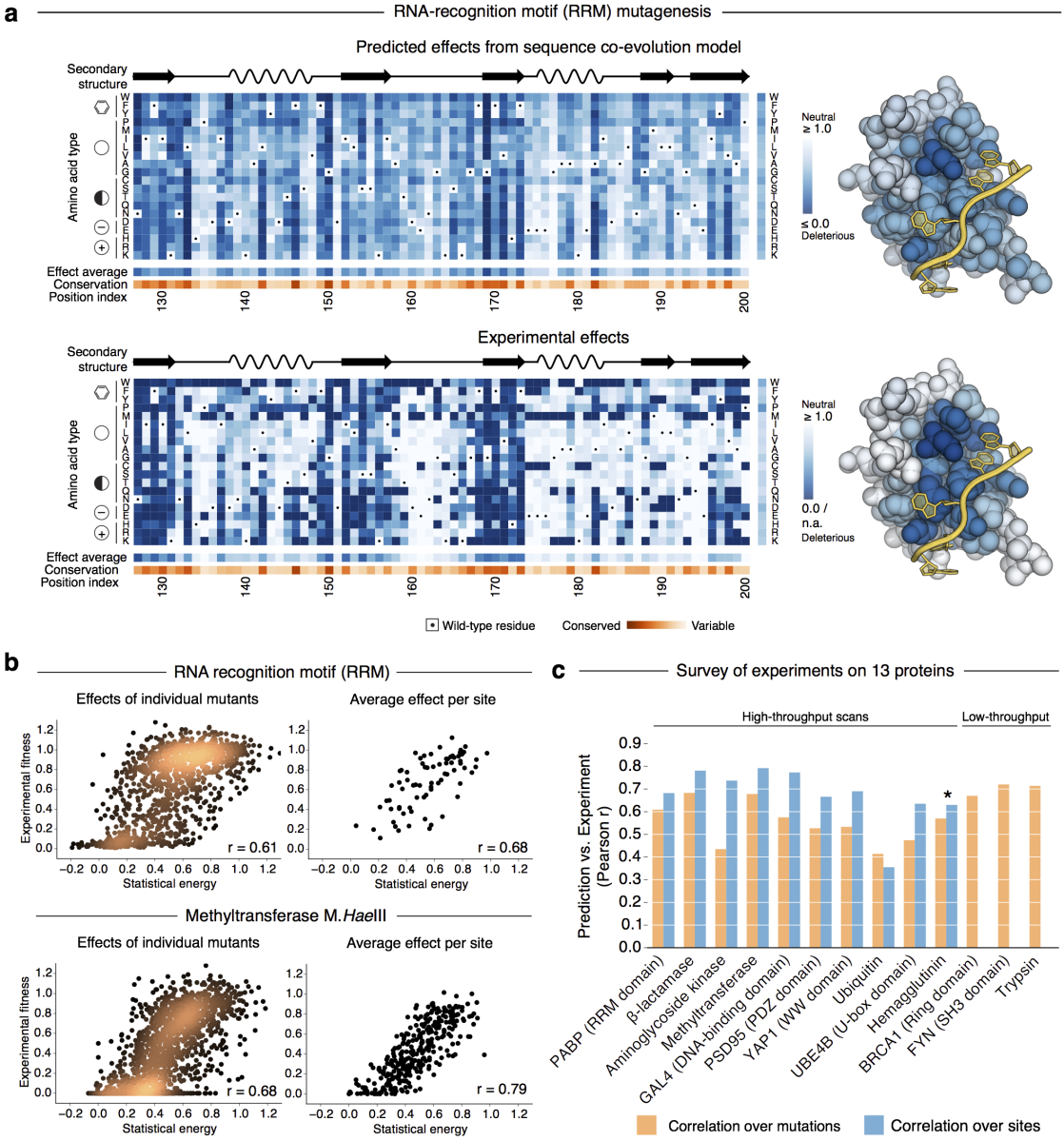

Figure 3.7.: Computed mutation effects agree with experimental phenotypic effects. (**a**) *Left*: Mutational landscape of single amino acid substitution effects in an RNA recognition motif (RRM) of the yeast poly(A)-binding protein computed from evolutionary sequence covariation (top panel) and tested experimentally *in vivo*[51] (bottom panel; x-axis: positions in RRM sequence, y-axis: amino acid substitutions, blue color: deleterious substitution, white color: neutral substitution). *Right:* Average mutational effect per position mapped on the 3D structure of human PABP (PDB: 1cvj[280]; RNA ligand in yellow). (**b**) Correlation between computed and experimental mutation effects in the RRM domain[51] (top) and the bacterial methyltransferase M.HaeIII[69] (bottom) for individual amino acid substitutions (left) and the average mutational effect per position (right). Some outliers on experimental axis are not shown in plots. (**c**) Correlation between mutation effects computed from sequence variation and experimental measurements (orange bars: individual mutations, blue bars: average effect per site) for all 13 tested proteins. (*) Hemagglutinin correlation was assessed on amino acids observed in sequence alignment only (3340 of 11280 substitutions). Adapted from Hopf *et al.*[3]

We found that the computed effects from evolutionary sequence variation correlate with experimental phenotypic effects for all of the 13 analyzed proteins, but with considerable spread in the strength of correlation (Figure 3.7a-c; Table A.2). For example, the model partially captures the effects on relative growth of 1179 point mutants of the RNA-recognition motif (RRM) of yeast polyadenylate-binding protein with a Pearson correlation of r=0.61; the average mutational sensitivity per site, i.e. the mean effect of all 19 possible substitutions, is predicted even more accurately (r=0.68; Figure 3.7a,b). Since the epistatic model can predict effects not only for single substitutions but also for higher-order mutations, we compared predictions and experiments for a set of 34745 RRM double mutants and obtained similar levels of correlation (r=0.62 for individual mutants; Figure A.1). Saturation mutagenesis has also been used to test the effects of single mutants on organismal growth of a bacterial DNA methylase, M.HaeIII from *Haemophilus aegyptius*, but using a considerably different experimental setup to dissect functional and non-functional variants in protecting DNA from restriction enzyme cleavage by multiple rounds of mutational drift[69]. Here, computed effects from sequence variation for 1685 single mutants correlate even stronger with the experimental fitness (r=0.68 for individual mutants, r=0.79 for site average; Figure 3.7b).

Overall, Pearson correlation coefficients across all analyzed proteins range from r=0.41 for a competitive growth experiment of 1270 yeast Ubiquitin single mutants[53] to r=0.72 for the *in vitro* protein stability (melting temperatures) of 47 single, double and triple mutants of the human FYN SH3 domain. Out of 13 analyzed experiments, 10 have correlation coefficients larger than 0.5, and for 6 r is larger than 0.6. For all but one of the deep mutational scanning experiments, the comparison over site averages gives correlations that are stronger than over individual mutants (r=0.36 to r=0.79), suggesting that averaging might reduce the influence of experimental noise and prediction errors (Figure 3.7c).

Since we visually observed non-linear relationships between predictions and experiments and bimodal effect distributions (Figure 3.7b, Figure A.2), which potentially violate the assumptions of Pearson's r, we additionally verified these results using: (i) the Spearman rank correlation coefficient ρ that tests for monotonic relationships between two variables and (ii) the Matthews correlation coefficient (MCC) to test binary classification accuracy (Section 2.4.2, Table A.2). For the same datasets, we obtained rank correlations from ρ=0.50 (Ubiquitin) to ρ=0.78 (rat anionic trypsin-2 stability[167]), largely confirming the results of the Pearson-based analysis with the exception of one outlier (homology-directed DNA repair function of 35 BRCA1 RING domain variants[70], r=0.67 vs. ρ=0.52) that was most likely caused by the bimodality of the experimental effect distribution. Similarly, the analysis in a binary classification setting based on the partitioning of experimental effects into deleterious and neutral mutations (Section 2.4.2) indicated that computed mutation effects are predictive of experimental phenotypes with MCCs ranging from 0.30 to 0.56 (when defining statistical

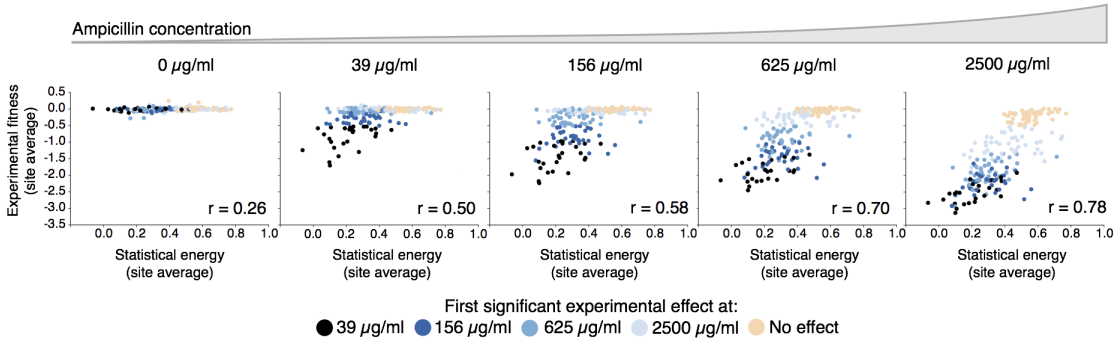

Figure 3.8.: Correlation between computed and experimental mutation effects depends on selective pressure. Amino acid substitutions (average mutational effect per site) of the β-lactamase TEM-1 have a deleterious computed effect (x-axis), but only show a deleterious effect *in vivo* as the selective pressure is increased using higher doses of the antibiotic ampicillin (left to right; concentrations of first significant experimental effect as determined by fitting a two-component Gaussian mixture model are highlighted by different shades of blue). Adapted from Hopf *et al.*[3]

energies <0.5 as deleterious predictions, similar results are obtained for thresholds of 0.4 and 0.6).

The results across the analyzed datasets, which are robust to different correlation measures, suggest that the computed effects from our epistatic model are reasonably predictive of experimentally tested phenotype changes upon mutation.

Agreement of predictions and experiments depends on experimental assay

The evolutionary information in sequence alignments of large evolutionary depth will usually be an aggregate over many selection experiments in different species living under different environmental conditions. Experiments on the other hand typically test the effects of mutations on a particular phenotype for one particular protein, and the assayed property may or may not be under selection. *A priori*, it is not clear if and how the signal from evolutionary sequence variation should be related to these phenotype changes. Experiments testing different phenotypes or environmental conditions for the same mutants however enable an assessment if particular features are under selection in the family, and could yield an explanation for the difference in the observed strengths of correlation.

Three of out the four analyzed experiments with the highest correlation for the average effect per site ($r \geq 0.7$) target bacterial proteins with a well-defined molecular function and corresponding selective pressures: the antibiotic resistance enzymes β-lactamase[71] and bacterial kinase APH(3')II[62], as well as the DNA methylase M.HaeIII that protects against DNA cleavage by the nuclease HaeIII[69]. For both antibiotic resistance enzymes, we observe a strong dependence of the correlations on the strength of purifying antibiotic selection applied during the experiments. Most mutations of β-lactamase are neutral to bacterial growth in the absence of its natural ligand ampi-

cillin (concentration 0 μg/ml) but predicted to be deleterious by our model (individual mutations: r=0.15, average effect per site: r=0.26). Only with increasing ampicillin concentrations and therefore dependence of bacterial growth on a functional enzyme, the correlation between experiments and computed effects becomes stronger (at 2500 μg/ml, individual mutations r=0.68 and average effect per site r=0.78; Figure 3.8). However, when selecting mutational variants using the non-natural, third-generation antibiotic cefotaxime, evolutionary information is not predictive of mutation effects anymore (r=-0.05 for individual mutations, r=-0.07 for average effect per site). This indicates that evolutionary information only captures mutation effects for the selective pressures the sequences evolved under.

For the bacterial kinase APH(3')II, which deactivates aminoglycoside antibiotics by phosphorylation, the highest correlations are observed for the lower tested concentrations of six antibiotics (1:8 or 1:4 dilution of the minimum inhibitory concentration (MIC) of the wild-type sequence), whereas effects saturate for the higher concentrations (1:2 and 1:1 MIC) and a large number of mutants is non-viable[62]. For example, under kanamycin selection correlations decrease from r=0.74 at 1:8 MIC to r=0.52 for 1:1 MIC (average effect per site); suggesting that the experimental pressure is too high at this point to discriminate between the relative strengths of different mutation effects.

We made observations similar to β-lactamase in a very different system that tested the effects of mutating a eukaryotic PDZ domain on peptide binding using a bacterial two-hybrid system[44]. Here, computed sequence information recapitulated the experimental mutation effects more accurately when measuring binding to the native ligand CRIPT (r=0.53, r=0.67 for average effect per site, 1600 mutations) than when testing binding to a non-native peptide (T-2F; r=0.31, r=0.36 for average effect per site). In the case of the PDZ domain, however, the coupling of mutation to phenotype under selective pressure is more challenging to assess experimentally than for the well-defined targets above, and the used artificial experimental system most likely can only capture a limited subset of the functional properties of this domain *in vivo*.

Some of the experimental studies analyzed in this work report measurements of multiple different phenotypes for each mutant, thus providing an opportunity to test which of these features might be under selection in the protein family. For example, a deep mutational scanning study of the RING domain of BRCA1 tested two aspects of molecular function of this protein, its E3 ligase activity and binding to the RING domain of BARD1[70]. On their own, these measurements correlate with predicted effects (r=0.38 and r=0.28, respectively), but not as strong as mutation effects on the overall molecular function of homology-directed DNA repair as inferred from a machine learning model combining the two orthogonal functional features (r=0.48)[70].

Two of the analyzed low-throughput, biochemical experiments individually measured the stability (melting temperatures) of mutants of rat anionic trypsin-2[167] and the SH3 domain of the human FYN oncogene[220], potentially giving more accurate results than high-throughput scans. The computational predictions correlate with the

melting temperatures more strongly than with any other of the analyzed datasets (Trypsin: r=0.71, SH3: r=0.72), suggesting that sufficient protein stability is an important factor in the evolution of these molecules[102,107]. This is in agreement with the observation that for both proteins, high-ranking evolutionary couplings correspond to residue contacts in their three-dimensional structures (precision of top N ECs > 0.9 at 8 Å distance threshold; Figure A.3, Table A.3). Correlations are however much lower for the catalytic activity of Trypsin ($\log k_{cat}/K_M$, r=-0.12) and the peptide-binding affinity of the SH3 domain ($\Delta\Delta G_{binding}$, r=-0.45). Although there is a possibility that the weaker correlation could be caused by a lack of selection for the particular phenotype, it seems more likely that the statistical model is unable to detect a signal here e.g. because the specificity for a particular ligand is only encoded in a small subfamily of the aligned sequences (lack of isofunctionality).

These results highlight that the interpretation of both mutation experiments and predictions, especially in an evolutionary context, critically depends on the particular phenotype that is assessed and the selective pressure to discriminate between neutral and deleterious genetic variants.

3.3.2. Contribution of epistatic interactions to predictions

The development of the epistatic model was motivated by the existing evidence for the context-dependence of genetic variants (Section 1.1.3). So far, we have established that computed mutation effects agree quantitatively and qualitatively with experimentally determined phenotypic consequences. To test if the accuracy of predictions benefits from the explicit incorporation of epistatic interactions, we constructed an equivalent first-order maximum entropy model that quantifies the probabilities of sequences using single-site terms h_i only (*independent model*, Section 2.3.3). We then inferred family-specific sets of parameters from the same sequence alignments as for the epistatic model, computed statistical energies for all experimentally tested mutations, and compared predictions versus experiments.

Epistatic model is a more accurate predictor of mutation effects

The predictions from the epistatic model on the analyzed set of 13 proteins are more accurate than those from the independent model as measured by Pearson correlations between prediction and experiment over individual mutations and average effects per site (Figure 3.9a, Table A.2). On the level of individual mutations, correlations for the epistatic model are substantially higher for 8 out of the 13 tested proteins and similar for the rest. When comparing average effects per site for the mutational scanning experiments, 7 out of 10 proteins have substantially higher correlations for the epistatic model. With the exception of the PDZ domain, where the epistatic model only outperforms the independent model on the average effect per site, epistatic model predictions are more accurate both on individual mutations and site averages. For none of the pro-

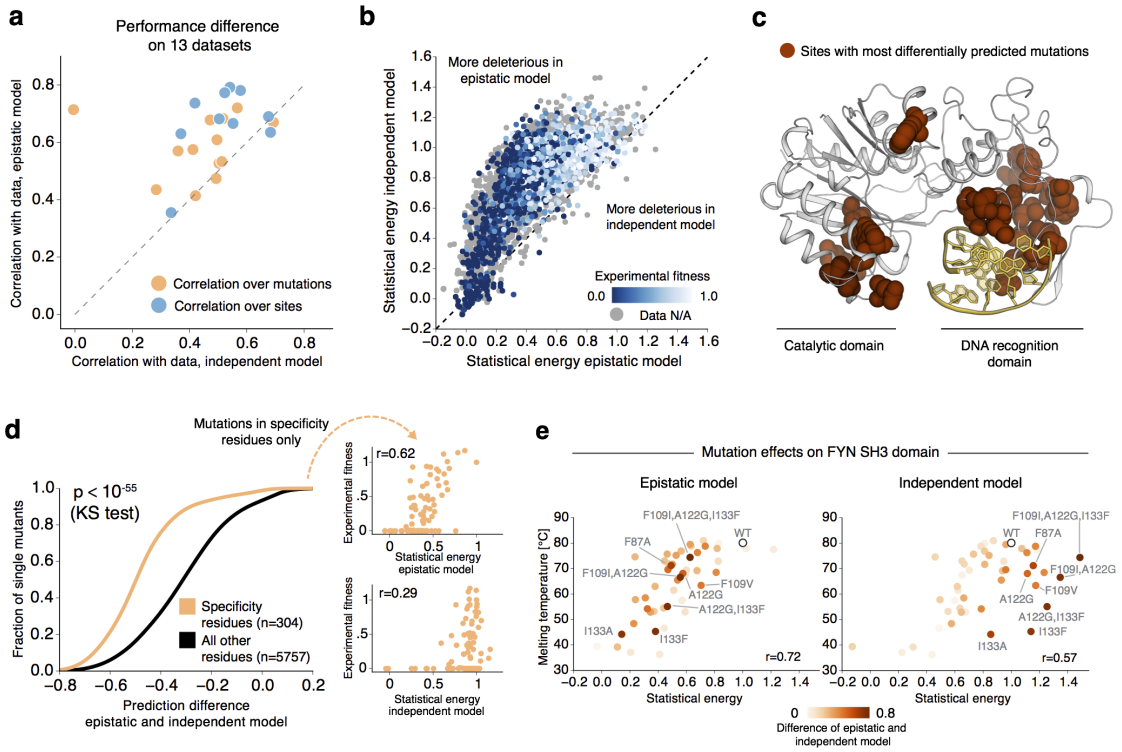

Figure 3.9.: Epistatic model predictions are more accurate than independent model on the tested set of 13 proteins. (**a**) Computed effects from the epistatic model correlate equally or more strongly with experimental measurements than effects from the independent model. (**b**) For the methyltransferase M.HaeIII, many computed mutation effects from the epistatic model are more deleterious than those from the independent model, which incorrectly predicts experimentally deleterious mutations (dark blue) as neutral. (**c**) Mutations that are most differentially predicted between the epistatic and independent models (red spheres, top 1% of single mutants) cluster in the catalytic and DNA recognition domains of M.HaeIII (PDB structure: 3ubt[281]). (**d**) *Left:* Computed effects for single mutations to specificity-determining residues in M.HaeIII (open conformation) are significantly more different between the epistatic and independent model than mutations to all other residues. *Right:* The effects from the epistatic model (top) correlate more strongly with experimental effects than the independent model (bottom) for specificity-determining residues. (**e**) The epistatic model is a more accurate predictor of melting temperatures in the human FYN SH3 domain than the independent model, which wrongly predicts multiple destabilizing mutations as neutral or beneficial. Adapted from Hopf *et al.*[3]

teins, the independent model performs substantially better. These results indicate that the epistatic model more accurately captures the amino acid constraints on the sequences in particular protein families. The correspondence of significant evolutionary couplings to three-dimensional structure contacts (Figure A.3) suggests that many of these epistatic dependencies are caused by the requirement to maintain a stable protein structure. It is important to note that the family-specific models are learned in an unsupervised setting from sequence data alone, without any reference to the experimental mutation effects. The improved predictive accuracy therefore cannot be caused

by overfitting to the experimental data using a higher number of parameters in the epistatic model compared to the independent model.

Improvement is pronounced for high effect and specificity-determining sites

A more detailed comparison of the predicted effects for individual mutations between both models revealed that predictions from the epistatic model tend to be more deleterious than those from the independent model, but not the other way round. Many of the neutral predictions under the independent model were however experimentally characterized as deleterious (Figure 3.9b, Figure A.4), suggesting that the epistatic model is a more accurate predictor for these mutants.

To test this hypothesis, we first analyzed the 5 proteins with deep mutational scanning data and the largest observed difference in predictive accuracy on individual mutations (β-lactamase, GAL4, bacterial kinase APH(3')II, M.HaeIII, PABP RRM domain; hemagglutinin was not analyzed here because of the large number of amino acids not evolutionarily observed). For all of these proteins except GAL4, differences in predicted effects between the epistatic and independent models were higher for mutations that were tested as experimentally deleterious than for the remaining mutations (Figure A.5a, Table A.4; statistical significance of difference in distributions assessed with two-sided sample Kolmogorov-Smirnov tests). Reversely, when evaluating predictive performance against experimental data on the subset of mutants with above-average prediction differences between the two models ($\Delta\Delta E^{\text{epi}}_{\text{ind}}(\boldsymbol{\sigma}^{(\text{mut})}, \boldsymbol{\sigma}^{(\text{wt})})$, Section 2.3.3 in Methods, Table A.5), the epistatic model gives substantially more accurate predictions than the independent model for the same 4 proteins. For the subset of mutants where both models give more similar predictions, i.e. with below-average prediction difference between, the correlations against the data agree as expected. Taken together, this suggests that the epistatic model is able to detect deleterious mutation effects that are hidden to the independent model; resulting in more accurate predictions overall.

Based on initial observations that mutations predicted the most differently by the epistatic and independent models tended to cluster around known functional regions of the proteins (Figure 3.9c), we systematically investigated if the use of epistatic interactions contributes to obtain more accurate predictions in these regions. For all of the 5 proteins above, detailed structural information about the ligand binding and protein interaction sites is available (β-lactamase, GAL4, bacterial kinase APH(3')II, M.HaeIII, PABP RRM domain). Except for β-lactamase with its conserved ligand, mutations to such specificity-determining sites (any residue within 4 Å of the ligand) are predicted more differently between the epistatic model and independent model than mutations to the other residues, and predictions from the epistatic model are more accurate when comparing against the experimental data (Figure A.5b, Table A.6). On the other hand, residues binding to conserved co-factors are predicted more similarly between both models than the remaining residues (GAL4, bacterial kinase APH(3')II, M.HaeIII). For example, specificity-determining residues in the open conformation of the DNA

methyltransferase M.HaeIII (PDB structure: 3ubt[281]) are predicted considerably more accurately by the epistatic model (r=0.62) than by the independent model (r=0.29, Figure 3.9d), whereas mutations to the sites binding the S-adenosylmethionine cofactor are predicted with comparable accuracy (r=0.74 vs. r=0.64, respectively). Together with the improved results on high-effect mutations, this indicates that the epistatic model is able to detect deleterious effects in sites that vary substantially between homologs, e.g. because of functional adaption, but are hidden to an independent model.

Structural interactions contribute to context-dependence of effects

The near-linear relationship between predicted statistical energies and stability measurements available for the FYN SH3 domain and Trypsin provides an opportunity to reliably identify which individual mutations are predicted more accurately by either model, i.e. deviate most strongly from the linear relationship (Figure 3.9e).

In the SH3 domain, mutant effects on thermostability are accurately predicted by the epistatic model (r=0.72) but less so by the independent model (r=0.57). The independent model incorrectly predicts several single and higher-order substitutions to residues F87, F109, A122 and I133 as neutral despite strong reductions in melting temperature of up to approximately 40°C (wild-type: 80.1°C). These four positions are proximal in structure in the core of the domain, and are connected by a network of high-scoring evolutionary couplings including the top-ranked pair between positions 122 and 133. Likewise, reductions in melting temperatures for mutants involving substitutions to residues M109 and C160 are not captured by the independent model; both residues have multiple strong evolutionary couplings to other structurally proximal positions.

In both cases, the deleterious substitutions are present frequently in homologs of the tested proteins, but not acceptable in the background of the wild-type sequence (cf. compensated pathogenic deviations, Section 1.1.3); here, the context-dependence appears to emerge from the coevolution of residues in structural contact. Only a context-dependent model will be able to explicitly identify such interactions and capture their influence on the phenotypic consequences of mutations.

Epistatic model outperforms existing effect prediction methods

While our assessment of the epistatic and independent models indicates that modeling pairwise interactions can improve the accuracy of mutation effect predictions, the majority of established computational methods does not consider the global sequence context (Section 1.2.2). To test how our sequence-based predictions compare to these approaches, we selected two representative state-of-the-art methods (SNAP2[114,115] and PolyPhen-2[120]) and predicted a high-confidence subset of seven of our analyzed datasets (RRM domain, β-lactamase, PDZ domain, methyltransferase M.HaeIII, GAL4, bacterial kinase APH(3')II, SH3 domain). It is important to note that

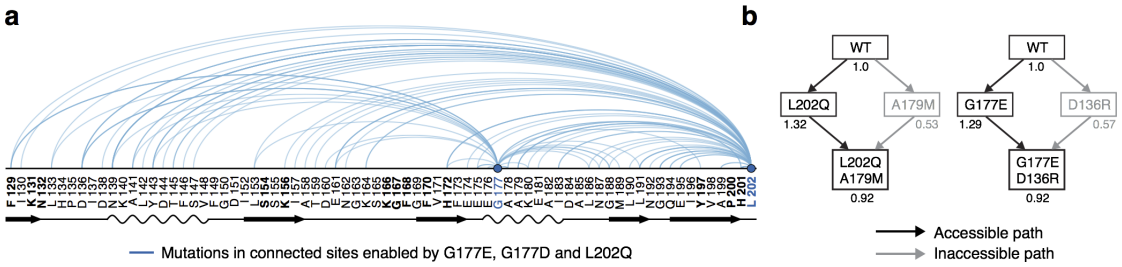

Figure 3.10.: Predicted permissive mutations in the RRM domain. (**a**) The substitutions G177D, G177E and L202Q are predicted permissive mutations enabling secondary substitutions in other sites (connected by blue arcs) that would be deleterious in the wild-type background (plot shows 100 most deleterious compensated mutations with statistical energy of double mutant $\geq$ 0.9, RNA-binding residues in bold font). (**b**) Examples of mutations that are quantified as deleterious on their own (A179M, D136R; statistical energy $<$ 0.9) but could occur as a secondary substitution after the permissive mutations L202Q and G177E. Adapted from Hopf *et al.*[3]

these methods are tailored towards the categorical classification of mutation effects, but also output quantitative scores that can be compared to quantitative mutation experiments. Both tested methods are based on machine learning on existing data and, besides sequence alignments, draw upon many additional input features such as structural information or database annotations.

On the tested proteins, the epistatic sequence model reaches similar or higher correlations against the data compared to both other methods (Figure A.6, Table A.7) despite only using sequence information. SNAP2 performs the most competitively to the epistatic model, while the PolyPhen-2 probabilistic classifier gives substantially less accurate predictions in a few cases. A major input feature to the PolyPhen-2 classifier is the PSIC conservation score[120], which on the level of individual mutations performs substantially worse than the epistatic model and approximately equally compared to the independent model. Taken together these observations suggest that (i) the epistatic model is a more accurate descriptor of mutation effects from sequence variation, and that (ii) machine learning-based methods, which improve upon the limited accuracy of single-site conservation using additional input features, could be enhanced by using an epistatic sequence model.

A refined comparative evaluation on more data and using extended metrics will however be necessary to validate these findings, which are somewhat dependent on the used correlation measure (Figure A.6, Table A.7). The tested methods tend to overpredict many mutations as deleterious, which could be caused by sequence alignments of limited evolutionary depth, and the interpretation of the continuous scores outside the anticipated binary classification use case may not be valid[114].

Epistatic model enables the prediction of mutational landscapes

A conceptual advance of the epistatic model compared to existing mutation effect prediction methods is its ability to compute the effects of higher-order mutations.

The evaluation of double and triple mutants of the RRM domain, Trypsin and the SH3 domain demonstrated that these experiments could be accurately predicted (Section 3.3.1). We therefore decided to apply our models for each of the analyzed proteins to a computational study of the mutational landscapes up to two substitutions away from the wild-type sequence of the target protein and the mutational paths connecting these sequences.

When defining a near wildtype-like statistical energy of at least 0.9 for any mutant to be considered as neutral or *viable*, only a very small fraction of double mutations passes this threshold (Table A.8). On average, only 0.4% of doubles are viable, with none of the proteins exceeding 1%. This suggests that the overwhelming majority of double mutants is deleterious, which is similar to figures reported for experimental higher-order mutations of a protein interaction and strongly limits the space of acceptable evolutionary trajectories[91].

Within this space of acceptable substitutions, on average 43% of doubles (range across proteins: 39%-81%) can be reached through both single mutants as viable intermediates (statistical energy ≥ 0.9), i.e. the order in which both singles occur does not matter. An average of 57% of double mutants (range: 19%-61%) can however only be reached through one of the single mutants, while the other intermediate is not accessible being non-viable (statistical energy < 0.9). In these cases, the first mutation enables the occurrence of the second mutation and therefore acts as a *permissive mutation* (Figure 3.10, Section 1.1.3). Several of our predicted permissive mutations have been described previously based on experimental studies, including G177E for the RRM domain[282] and N52A, R120G, L201P and M182T for β-lactamase[105,283,284].

While the exploration of mutational landscapes presented here is only an initial prototype study, it highlights the possibilities offered by this approach. Future work will need to systematically explore higher-order mutations, the pathways connecting them, test different viability thresholds and evaluate predictions against available experimental data.

3.3.3. Discussion

We have demonstrated that a probabilistic, epistatic model of evolutionary sequences can quantitatively predict the phenotypic consequences of amino acid substitutions in protein sequences. The explicit incorporation of all possible pairwise interactions between positions allows to obtain more accurate phenotype predictions than by assessing substitutions independently of the sequence background.

Coevolutionary models predict mutation effects

This work, which was inspired by the pioneering results of Lapedes *et al.*[139] on using pairwise maximum entropy models to predict residue contacts and the effects of mutations on stability, to the best of our knowledge represents one of the first two studies

to link sequence variation with large-scale mutational data while considering epistatic interactions. During the course of this project, independent work appeared that used the same or related formalisms to predict the effects of sequence changes on protein stability[143–145], viral fitness[140–142] and protein interactions[279]. These efforts confirm the promise of the probabilistic coevolutionary approach; however they were limited to small datasets on HIV variants[140–142] and some used three-dimensional structure information as part of the calculation[143–145]. An independent, parallel study by Figliuzzi *et al.*[285] with a main focus on predicting mutational landscapes for β-lactamase arrived at very similar conclusions to our work in that pairwise maximum entropy models quantify mutation effects more accurately than independent models and existing prediction methods.

Inference of epistatic models is statistically challenging

Despite the ability of epistatic models to predict mutation effects demonstrated here and elsewhere, their inference from sequences remains a difficult task. Currently, our model only considers pairwise interactions despite evidence for the importance of higher-order epistasis[286]. Even for pairwise interactions, the number of free parameters ($\binom{N}{2} \cdot (q-1)^2 + N \cdot (q-1)$ parameters for a protein of length N and $q = 21$ for 20 amino acids plus gap, e.g. approx. $2 \cdot 10^6$ for $N = 100$) exceeds the amount of available sequences (typically $< 10^5$) which makes parameter inference prone to overfitting[8,174,180]; this issue would aggravate for models that go beyond pairwise interactions, while their inference is in principle possible[174].

Although regularization is a basic strategy to reduce overfitting to limited training data, the inferred pair couplings J_{ij} still contain large amounts of bias that needs to be removed by the average product correction on the aggregated evolutionary couplings to obtain accurate contact predictions (Section 2.3.1)[171,175,180]. The statistical energy calculations used to predict mutation effects from individual parameters at present do not contain this correction, suggesting that better effect predictions could be obtained if the parameters were adjusted accordingly. Initial explorations based on scaling the regularization weight (variance of Gaussian priors) of individual position pairs, calculating subsets of the full statistical energy, or iterative learning of models to reduce the number of free parameters did however not lead to systematically better correlations against the data (data not shown).

More fundamentally, the observed sequences in protein families are of common evolutionary origin and usually have explored only a small fraction of the enormous space of all possible sequences (20^N for a protein of length N). The lack of training examples for most of sequence space will most likely lead to poor generalization to sequences that are too distant from those parts of sequence space that have been observed in evolution. This limitation calls for the development of quality metrics to blindly assess the reliability of statistical energies based on the shape of the observed parts of sequence space in the training data, and of models that address the biased nature of the data.

Relationship between different phenotypes and fitness remains elusive

A priori, it is not clear which exact relationship should be expected between our computed statistical energies and experimental measurements of *in vitro* and *in vivo* phenotypes, including growth as a proxy for fitness. The complex interplay between genotype and phenotypes on different levels of organization is still poorly understood, and to date only few studies have established quantitative links between them on selected examples[16,49,59,71,103,107,287–290]. A recurring theme is that changes in molecular phenotypes such as protein stability or catalytic activity have a non-linear influence on the viability of the organism, raising the question what traces selection leaves on the genotype level[16]. Not surprisingly, experiments that tested a well-defined selective pressure that maps directly to bacterial replication were amongst those that correlated the best with our computational predictions. In more complex cases, it seems unlikely that one experimental assay alone will capture the entirety of functional requirements on a protein. An evolutionary approach like the one presented here could help to assess the relevance of experimental assays; interpretations of such relationships are however dependent on the isofunctionality of the sequences in the alignment and their exposure to similar evolutionary forces.

Mutational scanning experiments are limited in resolution and interpretability

The assessment of the relationship between computational predictions and mutational scanning experiments is further complicated by several existing limitations of the current experimental techniques. First, the dynamic range of these assays is typically limited, leading to a loss of resolution for variants that are either very deleterious or have small effects[59]. This is directly evident for the analyzed studies that test the antibiotic resistance of β-lactamase and bacterial kinase mutants at different antibiotic concentrations, where sharp transitions occur between the effects measured at different pressures[59,71]; similar saturation effects can be observed when correlating the mutation effect measurements of a PDZ domain on binding to a native and a non-native ligand with lower binding affinity[44]. The presence of stability thresholds, where the ensemble for each mutant is either largely folded or unfolded, can further contribute to the occurrence of bimodal effect distributions and potentially creates additional dependencies on the exact environmental conditions of the experiment[63,87,94,103].

Second, results may be distorted by the use of artificial experimental systems to measure the functionality of the different variants. Several of the experiments analyzed here use phage display[37], two-hybrid systems[44] or special truncated sequence constructs[51] to select between functional and non-functional mutants, which may not be representative of the functional constraints *in vivo*[59]. Construction of such native assays is challenging in most cases, thus rendering antibiotic selection experiments a prime experimental system[59,62,71].

Third, high-throughput mutation experiments are subject to considerable noise and stochasticity[59]. Where available, biological replicates of the same experiment or independent studies of the same protein indicate substantial variation between measurements on the level of individual mutants. For example, the Pearson correlations between two biological replicates of a mutational scan of hemagglutinin are as low as r=0.54[66], and only r=0.48 when comparing to a different study[291]. Similarly, replicates for the bacterial kinase APH(3')II using the same antibiotic and concentration often do not exceed r=0.7, and correlations between experiments at different concentrations are even lower[62]. Due to the depletion of non-functional variants and low sequencing counts when testing many different variants at the same time, deleterious mutations are particularly susceptible to these effects[53,59,62,66] while the commonly used log-enrichment ratios emphasize the magnitude of their values.

These experimental limitations put an upper bound on the accuracy that can be expected when comparing computational predictions to them, and they will modulate the form of tested genotype-phenotype relationships in addition to the non-trivial dependencies between phenotypes on different levels as discussed above. Future work on experimental mutational scanning techniques and their interpretation is therefore critical to the assessment and development of methods as the one presented here.

Intra-molecular epistasis is prevalent within protein-coding sequences

The improved accuracy of phenotype predictions from the epistatic model compared to the independent model supports the context-dependence of the effects of amino acid substitutions in proteins, which has been a subject of intensive debate[90,95–98]. Even when considering the amino acid propensities of individual positions on their own, the observed amino acids contain implicit information about the functional requirements imposed by the remaining positions in the protein, and therefore about epistasis. Mutation effects can be highly dependent on their environment, but still be accurately predicted by an independent model. However, the more the remaining sequence backgrounds of the aligned sequences differ, the more this signal may get blurred. This hypothesis is in agreement with the tendency of the independent model to predict a subset of mutations as less deleterious than the epistatic model, e.g. experimentally deleterious substitutions as neutral (Section 3.3.2). If those predictions with explicit epistatic interaction are indeed more accurate when compared to experiments, as demonstrated here, then this can be interpreted as strong evidence for the abundant context-dependence of amino acid preferences and mutation effects in proteins.

Future research challenges

We have demonstrated that the probabilistic modeling of sequences with explicit consideration of epistatic interactions provides a reasonably accurate description of the specific sequence constraints in protein families. Subject to sufficient computational

resources and evolutionary sampling of sequence space, our approach allows to survey large numbers of sequence variants that are out of reach for experimental techniques. The current approach directly enables the assessment of genetic variation and exploration of protein evolutionary landscapes; some prototypes for such studies have been outlined here. Besides single proteins, the approach can be easily extended to other biomolecules such as RNA, and interactions between them (e.g. protein-protein, or protein-RNA)[279,292]. We have preliminary results suggesting that statistical energies can discriminate between functional and non-functional interactions of a bacterial two-component signaling complex (data not shown)[91].

A major prospect of our method is the application to protein design, e.g. for the de-immunization of molecules while retaining their function[139,293]. It will however be challenging to extrapolate from the observed samples under particular functional requirements to the design of novel functions.

3.4. Discussion of coevolution methods for phenotype prediction

So far, we have demonstrated that accurate predictions of protein structures, complexes and mutation effects can be obtained from the analysis of evolutionary sequence covariation. In this section, we discuss the general implications of our results across the different predicted phenotypes as well as the work published by others, and highlight challenges that will need to be addressed for the development of improved methods.

3.4.1. Implications of this and related work

After the proposition of coevolution methods in the 1990s, the last years have finally seen great progress in the accurate inference of coevolving residue pairs from alignments of evolutionary sequences (Section 1.2.3). Besides their practical applicability to relevant prediction problems, the success of these methods has profound implications on the computational analysis of sequence information.

Coevolution analysis reveals protein features hidden to single-site conservation

A trivial, yet fundamental contribution of coevolution methods is that they consider explicit interactions between different positions in sequences. The analysis of conservation patterns historically has been one of the pillars of computational biology, but the sequence context has been largely ignored or only considered in an implicit way[11,118,119,121]. A major reason for the lack of context-dependent analyses is the difficulty of the problem, which requires the application of sophisticated statistical models and sufficient data samples to infer pairwise or higher-order models[4,139,169,170,178]. With the introduction of global approaches like the pairwise maximum entropy models and the significant growth of sequence databases, these prerequisites were finally met[4], opening a new window into the interpretation of sequence information. Through the calculation of evolutionary couplings, a variety of phenotypes defined by interactions of positions can now be read off sequences alignments that are invisible to single sites. These include structural contacts in proteins and RNA molecules[8,139,173–175,292,294], their three-dimensional structures[8,9,183–188], conformational changes[9,189,190], protein interactions and multimerization[1,191,273,279]; and many others such as pathways of signal transmission could in principle be deduced[4,168]. Similar successes were reported for different problems like the analysis of neuron populations[295], chromatin factor interactions[296], functional couplings between genes[297] or the description of transcription factor binding sites[298]; highlighting the power of analyzing interactions in biological systems.

Protein biochemistry connects genotype to higher-order phenotypes

Our predictions of protein structures, complexes and mutation effects are based on the probabilistic modeling of sequence families. By simply interpreting the parameters of the models in different ways, such diverse phenotypes as 3D structure or overall effects of mutations on organismal growth can be predicted from patterns of amino acid covariation on the genotype level. The existence of this quantitative link from genotype to high-level phenotypes through structural intermediates suggests a critical and abundant role of intra- and inter-molecular epistasis in shaping the effects of mutations, and therefore of residue coevolution to maintain fitness. Ultimately, the biophysical interactions between different residues determine if proteins and their complexes are functional, and the selection for these context-dependent features is mirrored in the sequences[16]. Given the complexity of currently known mappings between genotype, biochemical phenotypes and fitness (Section 3.3.3), it seems however surprising that relatively simple probabilistic formalisms like the one used here are able to uncover these relationships at least in part.

3.4.2. Research challenges and future developments

Despite the considerable advances in interpreting the evolutionary record through the lens of coevolution, the outlined work only presents a starting point for future developments. Major challenges still need to be addressed, including the generation of sequence alignments with the right evolutionary depth, the interpretation of covariation patterns with regard to different functional features and phenotypes, and the application of the methods to uncharted examples and problems.

Choosing alignments of appropriate evolutionary scope

The raw material for the inference of conservation and covariation patterns on the genotype level is a sequence alignment for which one assumes isofunctionality to a conserved phenotypic property of interest, e.g. three-dimensional structure (Section 1.2). Alignments that contain too many sequences violating this assumption may lead to a loss of specificity; similarly, alignments that do not contain enough diverse examples may fail to create a signal in the first place or lead to wrong assessments based on the limited scope of observed acceptable substitutions[18,146]. A key challenge – that in principle affects any conservation-based method – is therefore to develop metrics that allow to blindly gauge the isofunctionality of sequences. This task is relatively easy and well-studied for individual protein structures and domains[198], but less so for many others such as protein interactions[299] or mutation effects (Section 1.2.2)[18,146]. Additionally, the optimal evolutionary scope might vary between different related features for the same protein. For example, overall structure could be conserved over a wide range of sequences while specific conformational changes upon ligand-binding might be limited to small subfamilies. We anticipate that computational approaches

can be devised to address this issue, e.g. through the iterative calculation of statistical energies to partition sequences into different functional subgroups and then updating the model parameters until convergence, or through the inference of mixture models that accommodate the existence of multiple subfamilies[300].

Besides measures ensuring sufficient isofunctionality of evolutionary sequences, the synthetic generation of sequences that are compatible with well-defined selective pressures under controlled environmental conditions could provide an alternative solution to obtain sufficient amounts of isofunctional genotype data[4,161]. It remains however unclear to what extent such experiments would need to explore sequence space to identify epistatic constraints in a comparable way to the myriads of mutation-selection experiments recorded in evolutionary sequence variation.

Disentangling signals for different functional features

The coevolutionary signal extracted from any sequence alignment will usually be a mixture of the different functional requirements on the aligned proteins. The development of methods to systematically disentangle which evolutionary couplings are caused by which particular feature remains an open challenge, e.g. for distinguishing between intra- and inter-subunit contacts in homomultimers or between different conformations of flexible proteins. Currently these problems tend to be solved using heuristic strategies, using additional information from orthogonal methods such as secondary structure or topology predictions (Section 2.3.2) or experimental structures for discrimination[274]. To eliminate the dependence on potentially error-prone predictions or the necessity for experimental information, future work ideally will test if the selection for particular functional features leaves specific evolutionary signatures that can be exploited to blindly interpret the pair couplings.

Possible applications and developments

Besides the challenges discussed so far, we anticipate that coevolutionary methods and evolutionary couplings could contribute particularly to two major fields of application, the engineering of biomolecules and the development of hybrid methods. Similar to how pair couplings made the problem of three-dimensional structure prediction tractable for many proteins that were out of reach of existing methods, the information about functionally critical interactions could focus the solution space of protein design efforts. For example, probabilistic approaches for sequence modeling could be used to find sequences that are compatible with the functional requirements of the wild-type sequence, but are not immunogenic by avoiding particular peptide sequences[293].

A second promising avenue is the development of hybrid methods that integrate computational predictions and experimental data to uncover phenotypes of interest. In a first example, we have developed an approach that combines evolutionary couplings and NMR measurements to obtain three-dimensional structures of proteins; in this

case NMR experiments provide detailed structural constraints on the local geometry of the peptide backbone while evolutionary couplings constrain the overall fold of the protein[6]. In combination, the orthogonal information from each method mutually compensates for the limitations of the other approach, giving access to results that would be much harder to obtain from the individual methods. Similar solutions could be devised for X-ray structure determination, where computational structure models could help to solve the molecular replacement problem, or for the detailed structural analysis of protein-protein interactions and their specificities[4]. Hybrid approaches are however not limited to structural phenotypes. For example, the combination of deep mutational scanning and evolution-based predictions of mutation effects could help to interpret the selective pressures acting on sequences *in vivo*, and guide the iterative design of new sequences that are refined experimentally starting from a computational exploration of sequence space.

Independent of the particular application, it will be crucial to develop better statistical models of sequence coevolution, more efficient methods for their inference, and to address the sequence alignment issues discussed above to allow the analysis of phenotypes on a genome-wide scale.

4. Conclusion

In this work, we have explored the use of probabilistic coevolutionary models of sequences to predict phenotypes from abundantly available protein-coding genotype data. The biophysical interactions between different residues lead to the context-dependence, or epistasis, of the phenotypic effects of amino acid substitutions[16]. By inferring patterns of amino acid conservation and covariation in sequence alignments using pairwise undirected graphical models, we were able to uncover these constraints on sequences left by evolutionary selection for particular phenotypes. Our analyses demonstrated that evolutionarily coupled positions in sequences frequently correspond to residue pair contacts in both monomer proteins and protein complexes and that this information can be used to reconstruct their three-dimensional structures. To confirm the validity of our approaches, we evaluated predictions on proteins and complexes that had experimental phenotype data, and found that the methods overall delivered accurate results when enough diverse sequences were available. Motivated by our earlier successes in the prediction of experimentally unsolved proteins, we computed structural models for proteins and complexes of interest, including insect olfactory receptors and the bacterial ATP synthase complex.

Besides these intermediate molecular phenotypes which are predicted based on epistatic constraints between positions, the family-specific changes in the probabilities of sequences correspond to experimentally tested phenotypic consequences of mutations, suggesting that the evolutionarily observed sampling of sequence space can be mapped to the phenotypes and fitness of synthetic sequences that have not been observed yet. The incorporation of epistatic interactions in these calculations improved the accuracy of the predicted effects compared to a model that treats sites independently, highlighting the context-dependence of acceptable amino acid substitutions. It will be an interesting challenge for future work to illuminate the relationship between molecular phenotypes, such as structure and complex formation, and the overall fitness of particular genotypes; and to identify the detailed molecular mechanisms by which epistasis arises.

Although many methodological challenges still need to be solved, the work presented here and related work by others highlights the promise of coevolutionary approaches to mine the wealth of available genotype data for phenotypic information. Going beyond the established concept of positional conservation, the lens of covariation offers another promising opportunity to interpret the beautiful experiment of evolution.

References

1. Hopf, T. A. *et al.* Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife* **3** (2014).
2. Hopf, T. A. *et al.* Amino acid coevolution reveals three-dimensional structure and functional domains of insect odorant receptors. *Nat Commun* **6,** 6077 (2015).
3. Hopf, T. A. *et al.* Quantification of the effect of mutations using a global probability model of natural sequence variation. arXiv: 1510.04612 (2015).
4. Marks, D. S., Hopf, T. A. & Sander, C. Protein structure prediction from sequence variation. *Nat. Biotechnol.* **30,** 1072–1080 (2012).
5. Kaján, L., Hopf, T. A., Kalaš, M., Marks, D. S. & Rost, B. FreeContact: fast and free software for protein contact prediction from residue co-evolution. *BMC Bioinformatics* **15,** 85 (2014).
6. Tang, Y. *et al.* Protein structure determination by combining sparse NMR data with evolutionary couplings. *Nat. Methods* **12,** 751–754 (2015).
7. Sheridan, R. *et al.* EVfold.org: Evolutionary Couplings and Protein 3D Structure Prediction. *bioRxiv,* 021022 (2015).
8. Marks, D. S. *et al.* Protein 3D structure computed from evolutionary sequence variation. *PLoS ONE* **6,** e28766 (2011).
9. Hopf, T. A. *et al.* Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* **149,** 1607–1621 (2012).
10. Hopf, T. & Kramer, S. in *Discovery Science* (eds Pfahringer, B., Holmes, G. & Hoffmann, A.) *Lecture Notes in Computer Science* 6332, 311–325 (Springer Berlin Heidelberg, 2010).
11. Radivojac, P. *et al.* A large-scale evaluation of computational protein function prediction. *Nat. Methods* **10,** 221–227 (2013).
12. Hamp, T. *et al.* Homology-based inference sets the bar high for protein function prediction. *BMC Bioinformatics* **14 Suppl 3,** S7 (2013).
13. Wagner, G. P. & Zhang, J. The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms. *Nat. Rev. Genet.* **12,** 204–213 (2011).
14. De Visser, J. A. G. M. & Krug, J. Empirical fitness landscapes and the predictability of evolution. *Nat. Rev. Genet.* **15,** 480–490 (2014).
15. Orgogozo, V., Morizot, B. & Martin, A. The differential view of genotype-phenotype relationships. *Front Genet* **6,** 179 (2015).

16. Harms, M. J. & Thornton, J. W. Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nat. Rev. Genet.* **14,** 559–571 (2013).

17. Lehner, B. Genotype to phenotype: lessons from model organisms for human genetics. *Nat. Rev. Genet.* **14,** 168–178 (2013).

18. Cooper, G. M. & Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat. Rev. Genet.* **12,** 628–640 (2011).

19. Lewontin, R. C. *et al. The genetic basis of evolutionary change* (Columbia University Press New York, 1974).

20. Smith, J. M. Natural selection and the concept of a protein space. *Nature* **225,** 563–564 (1970).

21. Whirl-Carrillo, M. *et al.* Pharmacogenomics knowledge for personalized medicine. *Clin. Pharmacol. Ther.* **92,** 414–417 (2012).

22. Dill, K. A., Ozkan, S. B., Shell, M. S. & Weikl, T. R. The protein folding problem. *Annu Rev Biophys* **37,** 289–316 (2008).

23. Benfey, P. N. & Mitchell-Olds, T. From genotype to phenotype: systems biology meets natural variation. *Science* **320,** 495–497 (2008).

24. Dowell, R. D. *et al.* Genotype to phenotype: a complex problem. *Science* **328,** 469 (2010).

25. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–753 (2009).

26. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucl. Acids Res.* **42,** D1001–D1006 (2014).

27. Giaever, G. *et al.* Functional profiling of the Saccharomyces cerevisiae genome. *Nature* **418,** 387–391 (2002).

28. Kim, D.-U. *et al.* Analysis of a genome-wide set of gene deletions in the fission yeast Schizosaccharomyces pombe. *Nat Biotech* **28,** 617–623 (2010).

29. Baba, T. *et al.* Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2** (2006).

30. Kamath, R. S. *et al.* Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature* **421,** 231–237 (2003).

31. Dietzl, G. *et al.* A genome-wide transgenic RNAi library for conditional gene inactivation in Drosophila. *Nature* **448,** 151–156 (2007).

32. Patwardhan, R. P. *et al.* High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat Biotech* **27,** 1173–1175 (2009).

33. Pitt, J. N. & Ferré-D'Amaré, A. R. Rapid construction of empirical RNA fitness landscapes. *Science* **330,** 376–379 (2010).

34. Sanjuán, R., Moya, A. & Elena, S. F. The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. *Proc. Natl. Acad. Sci. U.S.A.* **101,** 8396–8401 (2004).

35. Gold, M. G. *et al.* Molecular Basis of AKAP Specificity for PKA Regulatory Subunits. *Molecular Cell* **24,** 383–395 (2006).

36. Ernst, A. *et al.* Coevolution of PDZ domain-ligand interactions analyzed by high-throughput phage display and deep sequencing. *Mol Biosyst* **6,** 1782–1790 (2010).

37. Fowler, D. M. *et al.* High-resolution mapping of protein sequence-function relationships. *Nat. Methods* **7,** 741–746 (2010).

38. Hietpas, R. T., Jensen, J. D. & Bolon, D. N. A. Experimental illumination of a fitness landscape. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 7896–7901 (2011).

39. Hinkley, T. *et al.* A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase. *Nat Genet* **43,** 487–489 (2011).

40. Adkar, B. V. *et al.* Protein Model Discrimination Using Mutational Sensitivity Derived from Deep Sequencing. *Structure* **20,** 371–381 (2012).

41. Araya, C. L. *et al.* A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proc. Natl. Acad. Sci. U.S.A.* **109,** 16858–16863 (2012).

42. Deng, Z. *et al.* Deep sequencing of systematic combinatorial libraries reveals β-lactamase sequence constraints at high resolution. *J. Mol. Biol.* **424,** 150–167 (2012).

43. Fujino, Y. *et al.* Robust in vitro affinity maturation strategy based on interface-focused high-throughput mutational scanning. *Biochemical and Biophysical Research Communications* **428,** 395–400 (2012).

44. McLaughlin, R. N., Poelwijk, F. J., Raman, A., Gosal, W. S. & Ranganathan, R. The spatial architecture of protein function and adaptation. *Nature* **491,** 138–142 (2012).

45. Schlinkmann, K. M. *et al.* Critical features for biosynthesis, stability, and functionality of a G protein-coupled receptor uncovered by all-versus-all mutations. *Proc. Natl. Acad. Sci. U.S.A.* **109,** 9810–9815 (2012).

46. Traxlmayr, M. W. *et al.* Construction of a stability landscape of the CH3 domain of human IgG1 by combining directed evolution with high throughput sequencing. *J. Mol. Biol.* **423,** 397–412 (2012).

47. Whitehead, T. A. *et al.* Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nat. Biotechnol.* **30,** 543–548 (2012).

48. Forsyth, C. M. *et al.* Deep mutational scanning of an antibody against epidermal growth factor receptor using mammalian cell display and massively parallel pyrosequencing. *MAbs* **5,** 523–532 (2013).

49. Jiang, L., Mishra, P., Hietpas, R. T., Zeldovich, K. B. & Bolon, D. N. A. Latent effects of Hsp90 mutants revealed at reduced expression levels. *PLoS Genet.* **9,** e1003600 (2013).

50. Kim, I., Miller, C. R., Young, D. L. & Fields, S. High-throughput analysis of in vivo protein stability. *Mol. Cell Proteomics* **12,** 3370–3378 (2013).

51. 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).

52. Procko, E. *et al.* Computational Design of a Protein-Based Enzyme Inhibitor. *Journal of Molecular Biology* **425,** 3563–3575 (2013).

53. Roscoe, B. P., Thayer, K. M., Zeldovich, K. B., Fushman, D. & Bolon, D. N. A. Analyses of the effects of all ubiquitin point mutants on yeast growth rate. *J. Mol. Biol.* **425,** 1363–1377 (2013).

54. Starita, L. M. *et al.* Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. *Proc. Natl. Acad. Sci. U.S.A.* **110,** E1263–1272 (2013).

55. Tinberg, C. E. *et al.* Computational design of ligand-binding proteins with high affinity and selectivity. *Nature* **501,** 212–216 (2013).

56. Wu, N. C. *et al.* Systematic identification of H274Y compensatory mutations in influenza A virus neuraminidase by high-throughput screening. *J. Virol.* **87,** 1193–1199 (2013).

57. Acevedo, A., Brodsky, L. & Andino, R. Mutational and fitness landscapes of an RNA virus revealed through population sequencing. *Nature* **505,** 686–690 (2014).

58. Boucher, J. I. *et al.* Viewing protein fitness landscapes through a next-gen lens. *Genetics* **198,** 461–471 (2014).

59. Firnberg, E., Labonte, J. W., Gray, J. J. & Ostermeier, M. A comprehensive, high-resolution map of a gene's fitness landscape. *Mol. Biol. Evol.* **31,** 1581–1592 (2014).

60. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11,** 801–807 (2014).

61. Fowler, D. M., Stephany, J. J. & Fields, S. Measuring the activity of protein variants on a large scale using deep mutational scanning. *Nat. Protocols* **9,** 2267–2284 (2014).

62. Melnikov, A., Rogov, P., Wang, L., Gnirke, A. & Mikkelsen, T. S. Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. *Nucleic Acids Res.* **42,** e112 (2014).

63. 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).

64. Roscoe, B. P. & Bolon, D. N. A. Systematic exploration of ubiquitin sequence, E1 activation efficiency, and experimental fitness in yeast. *J. Mol. Biol.* **426,** 2854–2870 (2014).

65. Shin, H. *et al.* Exploring the functional residues in a flavin-binding fluorescent protein using deep mutational scanning. *PLoS ONE* **9,** e97817 (2014).

66. Thyagarajan, B. & Bloom, J. D. The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. *Elife* **3** (2014).

67. Wagenaar, T. R. *et al.* Resistance to vemurafenib resulting from a novel mutation in the BRAFV600E kinase domain. *Pigment Cell Melanoma Res* **27,** 124–133 (2014).

68. Kitzman, J. O., Starita, L. M., Lo, R. S., Fields, S. & Shendure, J. Massively parallel single-amino-acid mutagenesis. *Nat. Methods* **12,** 203–206 (2015).

69. Rockah-Shmuel, L., Tóth-Petróczy, Á. & Tawfik, D. S. Systematic Mapping of Protein Mutational Space by Prolonged Drift Reveals the Deleterious Effects of Seemingly Neutral Mutations. *PLoS Comput Biol* **11,** e1004421 (2015).

70. Starita, L. M. *et al.* Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. *Genetics* **200,** 413–422 (2015).

71. Stiffler, M. A., Hekstra, D. R. & Ranganathan, R. Evolvability as a function of purifying selection in TEM-1 β-lactamase. *Cell* **160,** 882–892 (2015).

72. Hietpas, R. T., Bank, C., Jensen, J. D. & Bolon, D. N. A. Shifting fitness landscapes in response to altered environments. *Evolution* **67,** 3512–3522 (2013).

73. Mercer, T. R. & Mattick, J. S. Structure and function of long noncoding RNAs in epigenetic regulation. *Nat Struct Mol Biol* **20,** 300–307 (2013).

74. Schuster, P., Fontana, W., Stadler, P. F. & Hofacker, I. L. From Sequences to Shapes and Back: A Case Study in RNA Secondary Structures. *Proc. R. Soc. Lond. B Biol. Sci.* **255,** 279–284 (1994).

75. Ferrada, E. & Wagner, A. A Comparison of Genotype-Phenotype Maps for RNA and Proteins. *Biophys J* **102,** 1916–1925 (2012).

76. Kendrew, J. C. *et al.* A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* **181,** 662–666 (1958).

77. Williamson, M. P., Havel, T. F. & Wüthrich, K. Solution conformation of proteinase inhibitor IIA from bull seminal plasma by 1H nuclear magnetic resonance and distance geometry. *J. Mol. Biol.* **182,** 295–315 (1985).

78. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28,** 235–242 (2000).

79. Dill, K. A. & MacCallum, J. L. The protein-folding problem, 50 years on. *Science* **338,** 1042–1046 (2012).

80. Moraes, I., Evans, G., Sanchez-Weatherby, J., Newstead, S. & Stewart, P. D. S. Membrane protein structure determination — The next generation. *Biochimica et Biophysica Acta (BBA) - Biomembranes. Structural and biophysical characterisation of membrane protein-ligand binding* **1838,** 78–87 (2014).

81. Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* **181,** 223–230 (1973).

82. Lehner, B. Molecular mechanisms of epistasis within and between genes. *Trends in Genetics* **27,** 323–331 (2011).

83. Costanzo, M. *et al.* The genetic landscape of a cell. *Science* **327,** 425–431 (2010).

84. Bloom, J. D., Gong, L. I. & Baltimore, D. Permissive secondary mutations enable the evolution of influenza oseltamivir resistance. *Science* **328,** 1272–1275 (2010).

85. Lunzer, M., Golding, G. B. & Dean, A. M. Pervasive cryptic epistasis in molecular evolution. *PLoS Genet.* **6,** e1001162 (2010).

86. Kondrashov, A. S., Sunyaev, S. & Kondrashov, F. A. Dobzhansky-Muller incompatibilities in protein evolution. *Proc. Natl. Acad. Sci. U.S.A.* **99,** 14878–14883 (2002).

87. DePristo, M. A., Weinreich, D. M. & Hartl, D. L. Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat. Rev. Genet.* **6,** 678–687 (2005).

88. Baresić, A., Hopcroft, L. E. M., Rogers, H. H., Hurst, J. M. & Martin, A. C. R. Compensated pathogenic deviations: analysis of structural effects. *J. Mol. Biol.* **396,** 19–30 (2010).

89. Jordan, D. M. *et al.* Identification of cis-suppression of human disease mutations by comparative genomics. *Nature* **524,** 225–229 (2015).

90. Breen, M. S., Kemena, C., Vlasov, P. K., Notredame, C. & Kondrashov, F. A. Epistasis as the primary factor in molecular evolution. *Nature* **490,** 535–538 (2012).

91. Podgornaia, A. I. & Laub, M. T. Protein evolution. Pervasive degeneracy and epistasis in a protein-protein interface. *Science* **347,** 673–677 (2015).

92. Capra, E. J. *et al.* Systematic dissection and trajectory-scanning mutagenesis of the molecular interface that ensures specificity of two-component signaling pathways. *PLoS Genet.* **6,** e1001220 (2010).

93. Skerker, J. M. *et al.* Rewiring the specificity of two-component signal transduction systems. *Cell* **133,** 1043–1054 (2008).

94. Bank, C., Hietpas, R. T., Jensen, J. D. & Bolon, D. N. A. A systematic survey of an intragenic epistatic landscape. *Mol. Biol. Evol.* **32,** 229–238 (2015).

95. McCandlish, D. M., Rajon, E., Shah, P., Ding, Y. & Plotkin, J. B. The role of epistasis in protein evolution. *Nature* **497,** E1–2, E1–2 (2013).

96. Pollock, D. D., Thiltgen, G. & Goldstein, R. A. Amino acid coevolution induces an evolutionary Stokes shift. *Proc. Natl. Acad. Sci. U.S.A.* **109,** E1352–1359 (2012).

97. Ashenberg, O., Gong, L. I. & Bloom, J. D. Mutational effects on stability are largely conserved during protein evolution. *Proc. Natl. Acad. Sci. U.S.A.* **110,** 21071–21076 (2013).

98. Pollock, D. D. & Goldstein, R. A. Strong evidence for protein epistasis, weak evidence against it. *Proc. Natl. Acad. Sci. U.S.A.* **111,** E1450 (2014).

99. Poon, A. & Chao, L. The rate of compensatory mutation in the DNA bacteriophage phiX174. *Genetics* **170,** 989–999 (2005).

100. Ivankov, D. N., Finkelstein, A. V. & Kondrashov, F. A. A structural perspective of compensatory evolution. *Curr. Opin. Struct. Biol.* **26,** 104–112 (2014).

101. Ortlund, E. A., Bridgham, J. T., Redinbo, M. R. & Thornton, J. W. Crystal structure of an ancient protein: evolution by conformational epistasis. *Science* **317,** 1544–1548 (2007).

102. Bloom, J. D., Labthavikul, S. T., Otey, C. R. & Arnold, F. H. Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. U.S.A.* **103,** 5869–5874 (2006).

103. Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N. & Tawfik, D. S. Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature* **444,** 929–932 (2006).

104. Tokuriki, N. & Tawfik, D. S. Stability effects of mutations and protein evolvability. *Curr. Opin. Struct. Biol.* **19,** 596–604 (2009).

105. Wang, X., Minasov, G. & Shoichet, B. K. Evolution of an Antibiotic Resistance Enzyme Constrained by Stability and Activity Trade-offs. *Journal of Molecular Biology* **320,** 85–95 (2002).

106. Weinreich, D. M., Delaney, N. F., Depristo, M. A. & Hartl, D. L. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* **312,** 111–114 (2006).

107. Gong, L. I., Suchard, M. A. & Bloom, J. D. Stability-mediated epistasis constrains the evolution of an influenza protein. *Elife* **2,** e00631 (2013).

108. Schenk, M. F., Szendro, I. G., Salverda, M. L. M., Krug, J. & de Visser, J. A. G. M. Patterns of Epistasis between beneficial mutations in an antibiotic resistance gene. *Mol. Biol. Evol.* **30,** 1779–1787 (2013).

109. Natarajan, C. *et al.* Epistasis among adaptive mutations in deer mouse hemoglobin. *Science* **340,** 1324–1327 (2013).

110. Alexander, P. A., He, Y., Chen, Y., Orban, J. & Bryan, P. N. A minimal sequence code for switching protein structure and function. *Proc. Natl. Acad. Sci. U.S.A.* **106,** 21149–21154 (2009).

111. Wylie, C. S. & Shakhnovich, E. I. A biophysical protein folding model accounts for most mutational fitness effects in viruses. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 9916–9921 (2011).

112. Jelier, R., Semple, J. I., Garcia-Verdugo, R. & Lehner, B. Predicting phenotypic variation in yeast from individual genome sequences. *Nat. Genet.* **43,** 1270–1274 (2011).

113. Grimm, D. G. *et al.* The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Hum. Mutat.* **36,** 513–523 (2015).

114. Hecht, M., Bromberg, Y. & Rost, B. Better prediction of functional effects for sequence variants. *BMC Genomics* **16 Suppl 8,** S1 (2015).

115. Bromberg, Y. & Rost, B. SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res.* **35,** 3823–3835 (2007).

116. Ng, P. C. & Henikoff, S. Predicting deleterious amino acid substitutions. *Genome Res.* **11,** 863–874 (2001).

117. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31,** 3812–3814 (2003).

118. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.* **39,** e118 (2011).

119. Reva, B., Antipin, Y. & Sander, C. Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biol.* **8,** R232 (2007).

120. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7,** 248–249 (2010).

121. Katsonis, P. & Lichtarge, O. A formal perturbation equation between genotype and phenotype determines the Evolutionary Action of protein-coding variations on fitness. *Genome Res.* **24,** 2050–2058 (2014).

122. Shihab, H. A. *et al.* Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum. Mutat.* **34,** 57–65 (2013).

123. Schwarz, J. M., Cooper, D. N., Schuelke, M. & Seelow, D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat. Methods* **11,** 361–362 (2014).

124. Chun, S. & Fay, J. C. Identification of deleterious mutations within three human genomes. *Genome Res.* **19,** 1553–1561 (2009).

125. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46,** 310–315 (2014).

126. Davydov, E. V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput. Biol.* **6,** e1001025 (2010).

127. Stone, E. A. & Sidow, A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res.* **15,** 978–986 (2005).

128. Capriotti, E., Calabrese, R. & Casadio, R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics* **22,** 2729–2734 (2006).

129. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15,** 1034–1050 (2005).

130. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20,** 110–121 (2010).

131. Asthana, S., Roytberg, M., Stamatoyannopoulos, J. & Sunyaev, S. Analysis of sequence conservation at nucleotide resolution. *PLoS Comput. Biol.* **3,** e254 (2007).

132. Gnad, F., Baucom, A., Mukhyala, K., Manning, G. & Zhang, Z. Assessment of computational methods for predicting the effects of missense mutations in human cancers. *BMC Genomics* **14 Suppl 3,** S7 (2013).

133. Potapov, V., Cohen, M. & Schreiber, G. Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. *Protein Eng. Des. Sel.* **22,** 553–560 (2009).

134. Guerois, R., Nielsen, J. E. & Serrano, L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* **320,** 369–387 (2002).

135. Rohl, C. A., Strauss, C. E. M., Misura, K. M. S. & Baker, D. Protein structure prediction using Rosetta. *Meth. Enzymol.* **383,** 66–93 (2004).

136. Schaefer, C., Bromberg, Y., Achten, D. & Rost, B. Disease-related mutations predicted to impact protein function. *BMC Genomics* **13,** S11 (2012).

137. Bromberg, Y., Kahn, P. C. & Rost, B. Neutral and weakly nonneutral sequence variants may define individuality. *Proc. Natl. Acad. Sci. U.S.A.* **110,** 14255–14260 (2013).

138. Hecht, M., Bromberg, Y. & Rost, B. News from the protein mutability landscape. *J. Mol. Biol.* **425,** 3937–3948 (2013).

139. Lapedes, A., Giraud, B. & Jarzynski, C. Using Sequence Alignments to Predict Protein Structure and Stability With High Accuracy. arXiv: 1207.2484 (2012).

140. Mann, J. K. *et al.* The fitness landscape of HIV-1 gag: advanced modeling approaches and validation of model predictions by in vitro testing. *PLoS Comput. Biol.* **10,** e1003776 (2014).

141. Ferguson, A. L. *et al.* Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design. *Immunity* **38,** 606–617 (2013).

142. Shekhar, K. *et al.* Spin models inferred from patient-derived viral sequence data faithfully describe HIV fitness landscapes. *Phys Rev E Stat Nonlin Soft Matter Phys* **88,** 062705 (2013).

143. Lui, S. & Tiana, G. The network of stabilizing contacts in proteins studied by coevolutionary data. *J Chem Phys* **139,** 155103 (2013).

144. Contini, A. & Tiana, G. A many-body term improves the accuracy of effective potentials based on protein coevolutionary data. *J Chem Phys* **143,** 025103 (2015).

145. Morcos, F., Schafer, N. P., Cheng, R. R., Onuchic, J. N. & Wolynes, P. G. Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. *Proc. Natl. Acad. Sci. U.S.A.* **111,** 12408–12413 (2014).

146. Hicks, S., Wheeler, D. A., Plon, S. E. & Kimmel, M. Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. *Hum. Mutat.* **32,** 661–668 (2011).

147. Chothia, C. & Lesk, A. M. The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5,** 823–826 (1986).

148. Sali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234,** 779–815 (1993).

149. Jacobson, M. & Sali, A. Comparative protein structure modeling and its applications to drug discovery. *Annu. Rep. Med. Chem.* **39,** 259–276 (2004).

150. Zhang, Q. C. *et al.* Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* **490,** 556–560 (2012).

151. Bradley, P., Misura, K. M. S. & Baker, D. Toward high-resolution de novo structure prediction for small proteins. *Science* **309,** 1868–1871 (2005).

152. De Vries, S. J. & Bonvin, A. M. J. J. CPORT: A Consensus Interface Predictor and Its Performance in Prediction-Driven Docking with HADDOCK. *PLoS ONE* **6,** e17695 (2011).

153. Hamp, T. & Rost, B. Evolutionary profiles improve protein–protein interaction prediction from sequence. *Bioinformatics* **31,** 1945–1950 (2015).

154. Dominguez, C., Boelens, R. & Bonvin, A. M. J. J. HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* **125,** 1731–1737 (2003).

155. Raman, S. *et al.* NMR structure determination for larger proteins using backbone-only data. *Science* **327,** 1014–1018 (2010).

156. 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).

157. Oosawa, K. & Simon, M. Analysis of mutations in the transmembrane region of the aspartate chemoreceptor in Escherichia coli. *PNAS* **83,** 6930–6934 (1986).

158. Neher, E. How frequent are correlated changes in families of protein sequences? *Proc. Natl. Acad. Sci. U.S.A.* **91,** 98–102 (1994).

159. Taylor, W. R. & Hatrick, K. Compensating changes in protein multiple sequence alignments. *Protein Eng.* **7,** 341–348 (1994).

160. Göbel, U., Sander, C., Schneider, R. & Valencia, A. Correlated mutations and residue contacts in proteins. *Proteins* **18,** 309–317 (1994).

161. 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).

162. Thomas, D. J., Casari, G. & Sander, C. The prediction of protein contacts from multiple sequence alignments. *Protein Eng.* **9,** 941–948 (1996).

163. Pazos, F., Helmer-Citterich, M., Ausiello, G. & Valencia, A. Correlated mutations contain information about protein-protein interaction. *J. Mol. Biol.* **271,** 511–523 (1997).

164. Pazos, F. & Valencia, A. Similarity of phylogenetic trees as indicator of protein-protein interaction. *Protein Eng.* **14,** 609–614 (2001).

165. Pazos, F. & Valencia, A. In silico two-hybrid system for the selection of physically interacting protein pairs. *Proteins* **47,** 219–227 (2002).

166. Lockless, S. W. & Ranganathan, R. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* **286,** 295–299 (1999).

167. Halabi, N., Rivoire, O., Leibler, S. & Ranganathan, R. Protein sectors: evolutionary units of three-dimensional structure. *Cell* **138,** 774–786 (2009).

168. Reynolds, K. A., McLaughlin, R. N. & Ranganathan, R. Hot spots for allosteric regulation on protein surfaces. *Cell* **147,** 1564–1575 (2011).

169. Lapedes, A. S., Giraud, B., Liu, L. & Stormo, G. D. in *Institute of Mathematical Statistics Lecture Notes - Monograph Series* 236–256 (Institute of Mathematical Statistics, Hayward, CA, 1999).

170. Giraud, B. G., Heumann, J. M. & Lapedes, A. S. Superadditive correlation. *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* **59,** 4983–4991 (1999).

171. 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).

172. De Juan, D., Pazos, F. & Valencia, A. Emerging methods in protein co-evolution. *Nat Rev Genet* **14,** 249–261 (2013).

173. 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. U.S.A.* **106,** 67–72 (2009).

174. Morcos, F. *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U.S.A.* **108,** E1293–1301 (2011).

175. Jones, D. T., Buchan, D. W. A., Cozzetto, D. & Pontil, M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **28,** 184–190 (2012).

176. Stein, R. R., Marks, D. S. & Sander, C. Inferring Pairwise Interactions from Biological Data Using Maximum-Entropy Probability Models. *PLoS Comput Biol* **11,** e1004182 (2015).

177. Burger, L. & van Nimwegen, E. Accurate prediction of protein–protein interactions from sequence alignments using a Bayesian method. *Mol. Syst. Biol.* **4** (2008).

178. Burger, L. & van Nimwegen, E. Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS Comput. Biol.* **6,** e1000633 (2010).

179. Aurell, E. & Ekeberg, M. Inverse Ising inference using all the data. *Phys. Rev. Lett.* **108,** 090201 (2012).

180. Ekeberg, M., Lövkvist, 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).

181. 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).

182. 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. U.S.A.* **110,** 15674–15679 (2013).

183. 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. U.S.A.* **109,** E1540–1547 (2012).

184. Sułkowska, J. I., Morcos, F., Weigt, M., Hwa, T. & Onuchic, J. N. Genomics-aided structure prediction. *PNAS* **109,** 10340–10345 (2012).

185. Kosciolek, T. & Jones, D. T. De novo structure prediction of globular proteins aided by sequence variation-derived contacts. *PLoS ONE* **9,** e92197 (2014).

186. Ovchinnikov, S. *et al.* Large scale determination of previously unsolved protein structures using evolutionary information. *eLife Sciences,* e09248 (2015).

187. Michel, M. *et al.* PconsFold: improved contact predictions improve protein models. *Bioinformatics* **30,** i482–488 (2014).

188. Hayat, S., Sander, C., Marks, D. S. & Elofsson, A. All-atom 3D structure prediction of transmembrane β-barrel proteins from sequences. *Proc. Natl. Acad. Sci. U.S.A.* **112,** 5413–5418 (2015).

189. Morcos, F., Jana, B., Hwa, T. & Onuchic, J. N. Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proc. Natl. Acad. Sci. U.S.A.* **110,** 20533–20538 (2013).

190. Sutto, L., Marsili, S., Valencia, A. & Gervasio, F. L. From residue coevolution to protein conformational ensembles and functional dynamics. *PNAS,* 201508584 (2015).

191. Schug, A., Weigt, M., Onuchic, J. N., Hwa, T. & Szurmant, H. High-resolution protein complexes from integrating genomic information with molecular simulation. *PNAS* **106,** 22124–22129 (2009).

192. Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7,** e1002195 (2011).

193. Remmert, M., Biegert, A., Hauser, A. & Söding, J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods* **9,** 173–175 (2012).

194. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).

195. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).

196. Durbin, R., Eddy, S. R., Krogh, A. & Mitchison, G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* (Cambridge University Press, 1998).

197. Povolotskaya, I. S. & Kondrashov, F. A. Sequence space and the ongoing expansion of the protein universe. *Nature* **465,** 922–926 (2010).

198. Sander, C. & Schneider, R. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins* **9,** 56–68 (1991).

199. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucl. Acids Res.* gkr367 (2011).

200. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7,** 539 (2011).

201. Suzek, B. E. *et al.* UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31,** 926–932 (2015).

202. UniProt Consortium. UniProt: a hub for protein information. *Nucleic Acids Res.* **43,** D204–212 (2015).

203. Pakseresht, N. *et al.* Assembly information services in the European Nucleotide Archive. *Nucleic Acids Res.* **42,** D38–43 (2014).

204. Besag, J. Statistical Analysis of Non-Lattice Data. *Journal of the Royal Statistical Society. Series D (The Statistician)* **24,** 179–195 (1975).

205. Obermayer, B. & Levine, E. Inverse Ising inference with correlated samples. *New J. Phys.* **16,** 123017 (2014).

206. Käll, L., Krogh, A. & Sonnhammer, E. L. L. An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics* **21 Suppl 1,** i251–257 (2005).

207. Nugent, T. & Jones, D. T. Transmembrane protein topology prediction using support vector machines. *BMC Bioinformatics* **10,** 159 (2009).

208. Bernsel, A., Viklund, H., Hennerdal, A. & Elofsson, A. TOPCONS: consensus prediction of membrane protein topology. *Nucleic Acids Res.* **37,** W465–468 (2009).

209. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292,** 195–202 (1999).

210. Brunger, A. T. Version 1.2 of the Crystallography and NMR system. *Nat Protoc* **2,** 2728–2733 (2007).

211. De Vries, S. J. *et al.* HADDOCK versus HADDOCK: new features and performance of HADDOCK2.0 on the CAPRI targets. *Proteins* **69,** 726–733 (2007).

212. Krivov, G. G., Shapovalov, M. V. & Dunbrack, R. L. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins* **77,** 778–795 (2009).

213. Sastry, G. M., Adzhigirey, M., Day, T., Annabhimoju, R. & Sherman, W. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J. Comput. Aided Mol. Des.* **27,** 221–234 (2013).

214. Tanabe, H. *et al.* Crystal structures of the human adiponectin receptors. *Nature* **520,** 312–316 (2015).

215. Baradaran, R., Berrisford, J. M., Minhas, G. S. & Sazanov, L. A. Crystal structure of the entire respiratory complex I. *Nature* **494,** 443–448 (2013).

216. Deng, D. *et al.* Crystal structure of the human glucose transporter GLUT1. *Nature* **510,** 121–125 (2014).

217. Jaremko, L., Jaremko, M., Giller, K., Becker, S. & Zweckstetter, M. Structure of the mitochondrial translocator protein in complex with a diagnostic ligand. *Science* **343,** 1363–1366 (2014).

218. Rajagopala, S. V. *et al.* The binary protein-protein interaction landscape of Escherichia coli. *Nat. Biotechnol.* **32,** 285–290 (2014).

219. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* **42,** D222–230 (2014).

220. Di Nardo, A. A., Larson, S. M. & Davidson, A. R. The relationship between conservation, thermodynamic stability, and function in the SH3 domain hydrophobic core. *J. Mol. Biol.* **333,** 641–655 (2003).

221. Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A* **32,** 922–923 (1976).

222. Lensink, M. F., Méndez, R. & Wodak, S. J. Docking and scoring protein complexes: CAPRI 3rd Edition. *Proteins* **69,** 704–718 (2007).

223. Méndez, R., Leplae, R., De Maria, L. & Wodak, S. J. Assessment of blind predictions of protein-protein interactions: current status of docking methods. *Proteins* **52,** 51–67 (2003).

224. Zhang, Y. & Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins* **57,** 702–710 (2004).

225. Xu, J. & Zhang, Y. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **26,** 889–895 (2010).

226. DeLano, W. L. The PyMOL molecular graphics system (2002).

227. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33,** 2302–2309 (2005).

228. Zou, K. H., Tuncali, K. & Silverman, S. G. Correlation and simple linear regression. *Radiology* **227,** 617–622 (2003).

229. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **405,** 442–451 (1975).

230. Oliphant, T. E. Python for Scientific Computing. *Comput. Sci. Eng.* **9,** 10–20 (2007).

231. Millman, K. J. & Aivazis, M. Python for Scientists and Engineers. *Comput. Sci. Eng.* **13,** 9–12 (2011).

232. Van der Walt, S., Colbert, S. & Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. *Comput. Sci. Eng.* **13,** 22–30 (2011).

233. Hunter, J. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9,** 90–95 (2007).

234. Seabold, S. & Perktold, J. *Statsmodels: Econometric and statistical modeling with python* in *Proceedings of the 9th Python in Science Conference* (2010), 57–61.

235. McKinney, W. *Data structures for statistical computing in Python* in *Proceedings of the 9th Python in Science Conference* **445** (2010), 51–56.

236. Pérez, F. & Granger, B. IPython: A System for Interactive Scientific Computing. *Comput. Sci. Eng.* **9,** 21–29 (2007).

237. Velankar, S. *et al.* SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Res.* **41,** D483–489 (2013).

238. Overington, J. P., Al-Lazikani, B. & Hopkins, A. L. How many drug targets are there? *Nat Rev Drug Discov* **5,** 993–996 (2006).

239. Moult, J., Pedersen, J. T., Judson, R. & Fidelis, K. A large-scale experiment to assess protein structure prediction methods. *Proteins* **23,** ii–v (1995).

240. Su, C.-Y., Menuz, K. & Carlson, J. R. Olfactory Perception: Receptors, Cells, and Circuits. *Cell* **139,** 45–59 (2009).

241. Benton, R. Multigene Family Evolution: Perspectives from Insect Chemoreceptors. *Trends in Ecology & Evolution* **30,** 590–600 (2015).

242. Biasini, M. *et al.* SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.* **42,** W252–258 (2014).

243. Benton, R. Chemical sensing in Drosophila. *Curr. Opin. Neurobiol.* **18,** 357–363 (2008).

244. Cherezov, V. *et al.* High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. *Science* **318,** 1258–1265 (2007).

245. Clyne, P. J. *et al.* A Novel Family of Divergent Seven-Transmembrane Proteins: Candidate Odorant Receptors in Drosophila. *Neuron* **22,** 327–338 (1999).

246. Hill, C. A. *et al.* G Protein-Coupled Receptors in Anopheles gambiae. *Science* **298,** 176–178 (2002).

247. Holm, L. & Rosenström, P. Dali server: conservation mapping in 3D. *Nucleic Acids Res.* **38,** W545–549 (2010).

248. Jin, X., Ha, T. S. & Smith, D. P. SNMP is a signaling component required for pheromone sensitivity in Drosophila. *Proc. Natl. Acad. Sci. U.S.A.* **105,** 10996–11001 (2008).

249. Leary, G. P. *et al.* Single mutation to a sex pheromone receptor provides adaptive specificity between closely related moth species. *Proc. Natl. Acad. Sci. U.S.A.* **109,** 14081–14086 (2012).

250. Pellegrino, M., Steinbach, N., Stensmyr, M. C., Hansson, B. S. & Vosshall, L. B. A natural polymorphism alters odour and DEET sensitivity in an insect odorant receptor. *Nature* **478,** 511–514 (2011).

251. Nichols, A. S. & Luetje, C. W. Transmembrane segment 3 of Drosophila melanogaster odorant receptor subunit 85b contributes to ligand-receptor interactions. *J. Biol. Chem.* **285,** 11854–11862 (2010).

252. Hughes, D. T., Wang, G., Zwiebel, L. J. & Luetje, C. W. A determinant of odorant specificity is located at the extracellular loop 2-transmembrane domain 4 interface of an Anopheles gambiae odorant receptor subunit. *Chem. Senses* **39,** 761–769 (2014).

253. Nakagawa, T., Pellegrino, M., Sato, K., Vosshall, L. B. & Touhara, K. Amino acid residues contributing to function of the heteromeric insect olfactory receptor complex. *PLoS ONE* **7,** e32372 (2012).

254. Wicher, D. *et al.* Drosophila odorant receptors are both ligand-gated and cyclic-nucleotide-activated cation channels. *Nature* **452,** 1007–1011 (2008).

255. Turner, R. M. *et al.* Mutational analysis of cysteine residues of the insect odorant co-receptor (Orco) from Drosophila melanogaster reveals differential effects on agonist- and odorant-tuning receptor-dependent activation. *J. Biol. Chem.* **289,** 31837–31845 (2014).

256. Kumar, B. N. *et al.* A conserved aspartic acid is important for agonist (VUAA1) and odorant/tuning receptor-dependent activation of the insect odorant co-receptor (Orco). *PLoS ONE* **8,** e70218 (2013).

257. Gofman, Y., Schärfe, C., Marks, D. S., Haliloglu, T. & Ben-Tal, N. Structure, dynamics and implied gating mechanism of a human cyclic nucleotide-gated channel. *PLoS Comput. Biol.* **10,** e1003976 (2014).

258. Wickles, S. *et al.* A structural model of the active ribosome-bound membrane protein insertase YidC. *Elife* **3,** e03035 (2014).

259. Abriata, L. A. Homology- and coevolution-consistent structural models of bacterial copper-tolerance protein CopM support a 'metal sponge' function and suggest regions for metal-dependent protein-protein interactions. *bioRxiv,* 013581 (2015).

260. Sato, K. *et al.* Insect olfactory receptors are heteromeric ligand-gated ion channels. *Nature* **452,** 1002–1006 (2008).

261. Liao, M., Cao, E., Julius, D. & Cheng, Y. Structure of the TRPV1 ion channel determined by electron cryo-microscopy. *Nature* **504,** 107–112 (2013).

262. Mosca, R., Céol, A. & Aloy, P. Interactome3D: adding structural details to protein networks. *Nat Meth* **10,** 47–53 (2013).

263. Johnson, E., Nguyen, P. T., Yeates, T. O. & Rees, D. C. Inward facing conformations of the MetNI methionine ABC transporter: Implications for the mechanism of transinhibition. *Protein Sci.* **21,** 84–96 (2012).

264. Kadaba, N. S., Kaiser, J. T., Johnson, E., Lee, A. & Rees, D. C. The High-Affinity E. coli Methionine ABC Transporter: Structure and Allosteric Regulation. *Science* **321,** 250–253 (2008).

265. Walker, J. E. The ATP synthase: the understood, the uncertain and the unknown. *Biochem. Soc. Trans.* **41,** 1–16 (2013).

266. Nguyen, P. T. *et al.* The contribution of methionine to the stability of the Escherichia coli MetNIQ ABC transporter-substrate binding protein complex. *Biol. Chem.* **396,** 1127–1134 (2015).

267. Ruangprasert, A. *et al.* Mechanisms of toxin inhibition and transcriptional repression by Escherichia coli DinJ-YafQ. *J. Biol. Chem.* **289,** 20559–20569 (2014).

268. Rastogi, V. K. & Girvin, M. E. Structural changes linked to proton translocation by subunit c of the ATP synthase. *Nature* **402,** 263–268 (1999).

269. Schwem, B. E. & Fillingame, R. H. Cross-linking between helices within subunit a of Escherichia coli ATP synthase defines the transmembrane packing of a four-helix bundle. *J. Biol. Chem.* **281,** 37861–37867 (2006).

270. Fillingame, R. H. & Steed, P. R. Half channels mediating H(+) transport and the mechanism of gating in the Fo sector of Escherichia coli F1Fo ATP synthase. *Biochim. Biophys. Acta* **1837,** 1063–1068 (2014).

271. Dmitriev, O., Jones, P. C., Jiang, W. & Fillingame, R. H. Structure of the membrane domain of subunit b of the Escherichia coli FoF1 ATP synthase. *J. Biol. Chem.* **274,** 15598–15604 (1999).

272. DeLeon-Rangel, J., Ishmukhametov, R. R., Jiang, W., Fillingame, R. H. & Vik, S. B. Interactions between subunits a and b in the rotary ATP synthase as determined by cross-linking. *FEBS Lett.* **587,** 892–897 (2013).

273. Ovchinnikov, S., Kamisetty, H. & Baker, D. Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. *eLife Sciences,* e02030 (2014).

274. Dos Santos, R. N., Morcos, F., Jana, B., Andricopulo, A. D. & Onuchic, J. N. Dimeric interactions and complex formation using direct coevolutionary couplings. *Sci. Rep.* **5,** 13652 (2015).

275. Linge, J. P., Habeck, M., Rieping, W. & Nilges, M. ARIA: automated NOE assignment and NMR structure calculation. *Bioinformatics* **19,** 315–316 (2003).

276. Nilges, M. A calculation strategy for the structure determination of symmetric dimers by 1H NMR. *Proteins* **17,** 297–309 (1993).

277. Thomas, J., Ramakrishnan, N. & Bailey-Kellogg, C. Graphical models of protein-protein interaction specificity from correlated mutations and interaction data. *Proteins* **76,** 911–929 (2009).

278. Kamisetty, H., Ghosh, B., Langmead, C. J. & Bailey-Kellogg, C. Learning sequence determinants of protein:protein interaction specificity with sparse graphical models. *J. Comput. Biol.* **22,** 474–486 (2015).

279. Cheng, R. R., Morcos, F., Levine, H. & Onuchic, J. N. Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. *Proc. Natl. Acad. Sci. U.S.A.* **111,** E563–571 (2014).

280. Deo, R. C., Bonanno, J. B., Sonenberg, N. & Burley, S. K. Recognition of polyadenylate RNA by the poly(A)-binding protein. *Cell* **98,** 835–845 (1999).

281. Didovyk, A. & Verdine, G. L. Structural origins of DNA target selection and nucleobase extrusion by a DNA cytosine methyltransferase. *J. Biol. Chem.* **287,** 40099–40105 (2012).

282. Melamed, D., Young, D. L., Miller, C. R. & Fields, S. Combining Natural Sequence Variation with High Throughput Mutational Data to Reveal Protein Interaction Sites. *PLoS Genet* **11,** e1004918 (2015).

283. Huang, W. & Palzkill, T. A natural polymorphism in β-lactamase is a global suppressor. *PNAS* **94,** 8801–8806 (1997).

284. Dellus-Gur, E., Toth-Petroczy, A., Elias, M. & Tawfik, D. S. What Makes a Protein Fold Amenable to Functional Innovation? Fold Polarity and Stability Trade-offs. *Journal of Molecular Biology* **425,** 2609–2621 (2013).

285. Figliuzzi, M., Jacquier, H., Schug, A., Tenaillon, O. & Weigt, M. Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase TEM-1. *Mol Biol Evol,* msv211 (2015).

286. Weinreich, D. M., Lan, Y., Wylie, C. S. & Heckendorn, R. B. Should evolutionary geneticists worry about higher-order epistasis? *Curr. Opin. Genet. Dev.* **23,** 700–707 (2013).

287. Bershtein, S., Mu, W. & Shakhnovich, E. I. Soluble oligomerization provides a beneficial fitness effect on destabilizing mutations. *PNAS* **109,** 4857–4862 (2012).

288. Bershtein, S., Mu, W., Serohijos, A. W. R., Zhou, J. & Shakhnovich, E. I. Protein Quality Control Acts on Folding Intermediates to Shape the Effects of Mutations on Organismal Fitness. *Molecular Cell* **49,** 133–144 (2013).

289. Serohijos, A. W. R. & Shakhnovich, E. I. Merging molecular mechanism and evolution: theory and computation at the interface of biophysics and evolutionary population genetics. *Curr. Opin. Struct. Biol.* **26,** 84–91 (2014).

290. Bershtein, S., Choi, J.-M., Bhattacharyya, S., Budnik, B. & Shakhnovich, E. Systems-Level Response to Point Mutations in a Core Metabolic Enzyme Modulates Genotype-Phenotype Relationship. *Cell Reports* **11,** 645–656 (2015).

291. Wu, N. C. *et al.* High-throughput profiling of influenza A virus hemagglutinin gene at single-nucleotide resolution. *Sci Rep* **4,** 4942 (2014).

292. Weinreb, C., Gross, T., Sander, C. & Marks, D. S. 3D RNA from evolutionary couplings. arXiv: 1510.01420 (2015).

293. Salvat, R. S., Parker, A. S., Choi, Y., Bailey-Kellogg, C. & Griswold, K. E. Mapping the Pareto Optimal Design Space for a Functionally Deimmunized Biotherapeutic Candidate. *PLoS Comput Biol* **11,** e1003988 (2015).

294. De Leonardis, E. *et al.* Direct-Coupling Analysis of nucleotide coevolution facilitates RNA secondary and tertiary structure prediction. *Nucl. Acids Res.* gkv932 (2015).

295. Schneidman, E., Berry, M. J., Segev, R. & Bialek, W. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* **440,** 1007–1012 (2006).

296. Zhou, J. & Troyanskaya, O. G. Global Quantitative Modeling of Chromatin Factor Interactions. *PLoS Comput Biol* **10,** e1003525 (2014).

297. Rivoire, O. Elements of Coevolution in Biological Sequences. *Phys. Rev. Lett.* **110,** 178102 (2013).

298. Santolini, M., Mora, T. & Hakim, V. A General Pairwise Interaction Model Provides an Accurate Description of In Vivo Transcription Factor Binding Sites. *PLoS ONE* **9,** e99015 (2014).

299. Mika, S. & Rost, B. Protein-protein interactions more conserved within species than across species. *PLoS Comput. Biol.* **2,** e79 (2006).

300. Ma, J., Wang, S., Wang, Z. & Xu, J. Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning. *Bioinformatics* (2015).

A. Supplementary Materials

Table A.1.: Dataset of experimental mutagenesis studies

Protein	UniProt ID	Assay	Phenotype	# Mutations	Comments	Ref.
TEM-1 β-lactamase	BLAT_ECOLX	Competitive growth under antibiotic selection	Growth	4997 (singles)	High correlation between 2 replicates; increasing concentrations of ampicillin and new ligand (cefotaxime)	71
TEM-1 β-lactamase	BLAT_ECOLX	Growth under antibiotic selection, band pass filter	Growth	5429 (singles)	Highly correlated to Stiffler *et al.* measurements	59
Polyadenylate-binding protein (RRM domain 2)	PABP_YEAST	Competitive growth	Growth	1188 (singles), 38352 (doubles)	Sequence tiled into 3 parts, doubles only within tiles; domain deletion in construct	51
PSD95 (PDZ domain 3)	DLG4_RAT	Bacterial two-hybrid (GFP expression), FACS selection	Peptide binding	1660 (singles)	Two-hybrid system only measures binding, no other functional requirements; 2 ligands measured	44
Modification methylase HaeIII	MTH3_HAEAE	Drift under selection	Growth	1957 (singles), 1738 (filtered)	Only mutants reachable by single nucleotide exchange, infrequent in initial library ($f \leq 0.01$) excluded from analysis	69
Aminoglycoside 3′-phosphotransferase	KKA2_KLEPN	Competitive growth under antibiotic selection	Growth	5280 (singles)	Excluded datapoints with very high fitness (> 5) which appear to be due to very low read counts	62
Regulatory protein GAL4 (DNA-binding domain)	GAL4_YEAST	Competitive growth on medium lacking histidine	Growth	1196 (singles)	–	68
Transcriptional coactivator YAP1 (WW domain 1)	YAP1_HUMAN	Phage display	Peptide binding	363 (singles), 9713 (doubles), 33166 (3-9 mut.)	Only analyzed single mutants	41
Ubiquitin	RL401_YEAST	Competitive growth	Growth	1270 (singles)	–	53
Ubiquitin	RL401_YEAST	Yeast display, fluorescent labeling of cells, FACS selection	E1 reactivity	1436 (singles)	–	64
Ubiquitination factor E4B (U-box domain)	UBE4B_MOUSE	Phage display, selection for UB ligase activity	Ligase activity	900 (singles), 90231 (2-9 mut.)	Only analyzed single mutants	54
Breast cancer type 1 susceptibility protein (RING domain)	BRCA1_HUMAN	Phage display (E3 reactivity), Y2H + reporter (BARD1 binding)	E3 ligase activity, BARD1 interaction, Homology-directed DNA repair (HDR)	4872 (E3), 1748 (BARD1), 44 (HDR)	Only 44 HDR experimental measurements, full matrix of HDR values predicted based on regression model	70
Hemagglutinin	(sequence from paper)	Passaging in tissue culture	Viral growth	11280 (singles)	Comparison to predictions only on substitutions that have been sampled by evolution in some background; transformed io log enrichment ratios	66
Tyrosine-protein kinase FYN (SH3 domain)	FYN_HUMAN	Low-throughput thermal denaturation and binding assays	Stability (T_m), peptide binding (K_d)	48 (WT, singles, doubles, triples)	Analyzed all datapoints with defined binding affinity; V138S excluded for non-cooperative unfolding	220
Anionic trypsin-2	TRY2_RAT	Low-throughput thermal denaturation and activity assays	Stability (T_m), catalytic activity (k_{cat}/K_m)	23 (WT, singles, doubles)	Hswap and C136A (no observable transition) were excluded from analysis	167

Table A.2.: Correlations between computed and experimental mutation effects

				Correlation over mutations						Correlation over sites			
				Pearson r		Spearman ρ		MCC		Pearson r		Spearman ρ	
Protein	UniProt ID	Ref.	Experiment	E[a]	I[b]	E	I	E	I	E	I	E	I
TEM-1 β-lactamase	BLAT_ECOLX	71	2500	0.68	0.52	0.71	0.52	0.50	0.28	0.78	0.58	0.78	0.60
BRCA1 (Ring domain)	BRCA1_HUMAN	70	hdr	0.67	0.69	0.52	0.50	–	–	–	–	–	–
PSD95 (PDZ domain)	DLG4_RAT	44	CRIPT	0.53	0.50	0.53	0.44	0.42	0.47	0.67	0.55	0.72	0.60
FYN (SH_3 domain)	FYN_HUMAN	220	Tm	0.72	0.57	0.75	0.51	–	–	–	–	–	–
GAL4 (DNA-binding domain)	GAL4_YEAST	68	SEL_C_40h	0.58	0.41	0.58	0.40	0.36	0.28	0.77	0.52	0.80	0.55
Hemagglutinin[c]	–	66	log_ratio	0.57	0.35	0.61	0.43	–	–	0.62	0.36	0.61	0.42
Bacterial kinase	KKA2_KLEPN	62	KKA2_S3_Kan18_L1 (filtered)	0.44	0.28	0.52	0.28	0.48	0.33	0.74	0.42	0.75	0.39
Methyltransferase	MTH3_HAEAE	69	17_filtered	0.68	0.47	0.69	0.44	0.55	0.33	0.79	0.54	0.81	0.48
PABP (RRM domain)	PABP_YEAST	51	linear	0.61	0.50	0.59	0.44	0.45	0.39	0.68	0.50	0.69	0.46
Ubiquitin	RL401_YEAST	53	selection_coefficient	0.41	0.42	0.50	0.45	0.30	0.29	0.36	0.34	0.43	0.47
Trypsin	TRY2_RAT	167	Tm	0.71	-0.01	0.78	0.00	–	–	–	–	–	–
UBE4B (U-box domain)	UBE4B_MOUSE	54	log2_ratio	0.47	0.49	0.52	0.49	0.39	0.35	0.64	0.68	0.65	0.66
YAP1 (WW domain)	YAP1_HUMAN	41	linear	0.53	0.51	0.60	0.58	0.56	0.42	0.69	0.68	0.75	0.78

[a] Epistatic model [b] Independent model [c] Evaluated only on amino acids observed in alignment

Table A.3.: Correspondence of evolutionary couplings to 3D residue contacts

	Precision top N couplings[a]			Precision significant couplings[b]		
UniProt ID	#	5 Å[c]	8 Å	#	5 Å	8 Å
BLAT_ECOLX	215	0.66	0.88	879	0.36	0.64
BRCA1_HUMAN	76	0.59	0.75	103	0.50	0.67
DLG4_RAT	80	0.66	0.84	146	0.47	0.67
FYN_HUMAN	53	0.79	0.91	95	0.57	0.82
GAL4_YEAST	63	0.51	0.76	81	0.46	0.69
KKA2_KLEPN	213	0.52	0.82	918	0.25	0.53
MTH3_HAEAE	319	0.57	0.78	1686	0.21	0.42
PABP_YEAST	76	0.82	0.99	221	0.49	0.77
RL401_YEAST	71	0.72	0.86	129	0.54	0.71
TRY2_RAT	217	0.85	0.99	1016	0.40	0.65
UBE4B_MOUSE	76	0.54	0.82	116	0.44	0.68
YAP1_HUMAN	31	0.71	0.77	34	0.74	0.79

[a] N=length of statistical model [b] Coupling scores above quality score threshold [c] Distance (minimum atom) threshold of experimental residue contact

Table A.4.: Prediction difference between epistatic and independent models on mutations with high experimental effect

		Number of mutations		
UniProt ID	Effect threshold[a]	Deleterious	Neutral	P-value[b]
BLAT_ECOLX	-0.34	2575	1510	$5.9 \cdot 10^{-32}$
DLG4_RAT	-0.28	333	1267	$1.1 \cdot 10^{-4}$
GAL4_YEAST	-7.22	363	760	0.10
KKA2_KLEPN	0.52	1572	2677	$1.3 \cdot 10^{-31}$
MTH3_HAEAE	0.46	895	790	$1.3 \cdot 10^{-14}$
PABP_YEAST	0.57	404	775	$7.1 \cdot 10^{-13}$
RL401_YEAST	-0.20	406	825	0.40
UBE4B_MOUSE	-0.57	314	300	0.40
YAP1_HUMAN	0.43	110	209	0.91

[a] Determined by fitting a two-component Gaussian mixture model to experimental effect distribution [b] Two-sided sample Kolmogorov-Smirnov test

Table A.5.: Correlations of mutation effects predicted similarly and differently between epistatic and independent models with experimental data

	Number of mutations		All mutations		Different		Similar	
UniProt ID	Different[a]	Similar[a]	E[b]	I[c]	E	I	E	I
BLAT_ECOLX	2089	1996	0.68	0.52	0.62	0.46	0.74	0.71
DLG4_RAT	739	861	0.53	0.50	0.49	0.41	0.59	0.61
GAL4_YEAST	535	588	0.57	0.41	0.31	0.30	0.69	0.65
KKA2_KLEPN	2036	2213	0.43	0.28	0.43	0.25	0.42	0.45
MTH3_HAEAE	824	861	0.68	0.47	0.52	0.32	0.75	0.72
PABP_YEAST	548	631	0.61	0.50	0.52	0.31	0.65	0.69
RL401_YEAST	627	604	0.41	0.42	0.35	0.32	0.50	0.53
UBE4B_MOUSE	286	328	0.47	0.49	0.41	0.31	0.58	0.62
YAP1_HUMAN	146	173	0.53	0.51	0.48	0.48	0.57	0.58

[a] Mutation effects more and less different than average $\Delta\Delta E^{\text{epi}}_{\text{ind}}(\boldsymbol{\sigma}^{(\text{mut})}, \boldsymbol{\sigma}^{(\text{wt})})$ [b] Epistatic model (Pearson r) [c] Independent model (Pearson r)

Table A.6.: Prediction difference between epistatic and independent models on specificity-determining sites

			Number of mutations			Correlation with data[c]	
UniProt ID	Functional feature	Positions[a]	These positions	Others	P-value[b]	E[d]	I[e]
BLAT_ECOLX	ligand (mostly conserved)	67, 68, 103, 128, 130, 214, 232, 233, 234, 235, 236, 241	228	3857	$1.1 \cdot 10^{-7}$	0.62	0.59
GAL4_YEAST	ligand specificity	8, 9, 10, 15, 17, 18, 19, 20, 21, 23, 43, 49, 50, 51	266	931	$6.2 \cdot 10^{-12}$	0.40	0.37
GAL4_YEAST	cofactor	11, 14, 21, 28, 31, 38	114	1083	$2.9 \cdot 10^{-59}$	0.27	0.19
KKA2_KLEPN	cofactor	27, 29, 32, 36, 47, 49, 94, 95, 96, 97, 194, 195, 197, 207, 208	285	3762	$1.3 \cdot 10^{-5}$	0.55	0.50
KKA2_KLEPN	ligand specificity	157, 158, 159, 160, 190, 211, 226, 227, 230	171	3876	$5.7 \cdot 10^{-07}$	0.49	0.19
MTH3_HAEAE	ligand specificity[f]	68, 69, 70, 71, 72, 75, 76, 77, 78, 79, 80, 81, 87, 109, 111, 112, 114, 116, 117, 118, 152, 153, 155, 198, 217, 219, 220, 221, 224, 225, 227, 229, 236, 237, 239, 240, 241, 243, 244, 260, 305, 306	798	5263	$3.8 \cdot 10^{-32}$	0.65	0.51
MTH3_HAEAE	cofactor	7, 8, 9, 10, 12, 13, 28, 29, 30, 31, 50, 51, 68, 70, 90, 286, 306, 307, 308	361	5700	$1.1 \cdot 10^{-23}$	0.74	0.64
PABP_YEAST	ligand specificity	127, 129, 131, 132, 154, 156, 166, 167, 168, 170, 172, 197, 200, 201, 202	285	1159	$3.5 \cdot 10^{-11}$	0.44	0.16
PABP_YEAST	protein interaction	137, 139, 140, 141, 145, 148, 149, 153, 154, 155, 156, 186, 188, 189, 190, 191, 192, 193	342	1102	$2.3 \cdot 10^{-18}$	0.52	0.23
PABP_YEAST	ligand and protein specificity	127, 129, 131, 132, 137, 139, 140, 141, 145, 148, 149, 153, 154, 154, 155, 156, 156, 166, 167, 168, 170, 172, 186, 188, 189, 190, 191, 192, 193, 197, 200, 201, 202	589	855	$1.6 \cdot 10^{-19}$	0.50	0.25

[a] Residues within 4 Å minimum atom distance of ligand/interaction partner [b] Two-sided sample Kolmogorov-Smirnov test [c] Pearson r [d] Epistatic model [e] Independent model [f] Merged ligand contacts in open and closed conformations, while Figure 3.9d is based on open conformation only.

Table A.7.: Comparison with machine learning-based prediction methods

			Correlation over mutations[a]		Correlation over sites[a]	
UniProt ID	Experiment	method	Pearson r	Spearman ρ	Pearson r	Spearman ρ
BLAT_ECOLX	2500	Epistatic model	0.68	0.71	0.78	0.78
		Independent model	0.52	0.51	0.58	0.60
		PolyPhen-2 (PSIC)	0.55	0.56	0.69	0.72
		PolyPhen-2 (Probability)	0.57	0.68	0.76	0.78
		SNAP2	0.68	0.71	0.74	0.74
PABP_YEAST	linear	Epistatic model	0.61	0.59	0.68	0.69
		Independent model	0.50	0.44	0.50	0.46
		PolyPhen-2 (PSIC)	0.39	0.43	0.56	0.55
		PolyPhen-2 (Probability)	0.41	0.46	0.55	0.59
		SNAP2	0.51	0.55	0.53	0.56
DLG4_RAT	CRIPT	Epistatic model	0.53	0.53	0.67	0.72
		Independent model	0.50	0.44	0.55	0.60
		PolyPhen-2 (PSIC)	0.25	0.25	0.45	0.53
		PolyPhen-2 (Probability)	0.22	0.34	0.39	0.54
		SNAP2	0.45	0.56	0.56	0.67
MTH3_HAEAE	17 (filtered)	Epistatic model	0.68	0.69	0.79	0.81
		Independent model	0.47	0.44	0.54	0.48
		PolyPhen-2 (PSIC)	0.59	0.60	0.73	0.74
		PolyPhen-2 (Probability)	0.65	0.69	0.81	0.83
		SNAP2	0.64	0.64	0.74	0.74
KKA2_KLEPN	KKA2_S3_Kan18_L1	Epistatic model	0.43	0.52	0.74	0.75
	(filtered)	Independent model	0.28	0.28	0.42	0.38
		PolyPhen-2 (PSIC)	0.37	0.41	0.65	0.68
		PolyPhen-2 (Probability)	0.38	0.51	0.71	0.74
		SNAP2	0.48	0.54	0.73	0.73
GAL4_YEAST	SEL_C_40h	Epistatic model	0.57	0.58	0.77	0.80
		Independent model	0.41	0.40	0.52	0.55
		PolyPhen-2 (PSIC)	0.37	0.36	0.75	0.70
		PolyPhen-2 (Probability)	0.50	0.54	0.74	0.66
		SNAP2	0.51	0.51	0.63	0.62
FYN_HUMAN	Tm	Epistatic model	0.72	0.73	–	–
		Independent model	0.61	0.57	–	–
		PolyPhen-2 (PSIC)	0.60	0.58	–	–
		PolyPhen-2 (Probability)	0.35	0.50	–	–
		SNAP2	0.52	0.56	–	–

[a] Absolute value of correlation coefficient

Table A.8.: Computed double mutant landscapes

				% of viable doubles accessible through[b]		
UniProt ID	# Double mutants	# Viable doubles[a]	% Viable[a]	0 paths	1 path	2 paths
PABP_YEAST	1028850	10505	1.02	0	46	54
BLAT_ECOLX	8304805	1728	0.02	0	47	53
KKA2_KLEPN	8150658	17821	0.22	0	57	43
MTH3_HAEAE	18310281	19758	0.11	0	50	50
GAL4_YEAST	705033	5429	0.77	0	39	61
DLG4_RAT	1140760	4404	0.39	0	53	47
YAP1_HUMAN	167865	344	0.20	0	44	56
RL401_YEAST	897085	75	0.01	0	81	19
UBE4B_MOUSE	1028850	3991	0.39	0	74	26
BRCA1_HUMAN	1028850	4593	0.45	0	78	22
FYN_HUMAN	497458	2911	0.59	0	60	40
TRY2_RAT	8460396	26419	0.31	0	61	39

[a]Viable mutants are those with statistical energy ≥ 0.9 [b]Accessible paths lead through single mutants with statistical energy ≥ 0.9

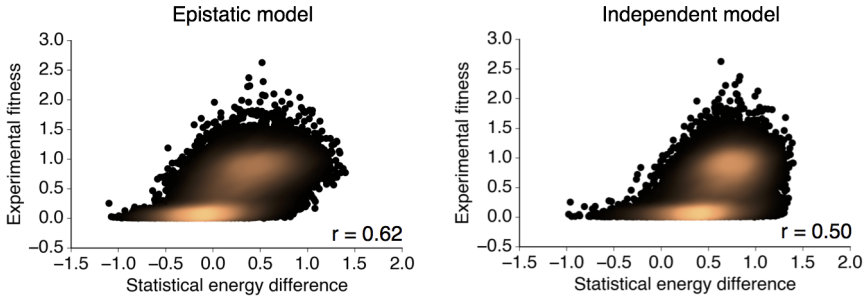

Figure A.1.: Double mutations in the PABP RRM domain. Mutation effects computed using the epistatic model agree more strongly with the experimental effects of 34745 double mutants in the RRM domain (r=0.62) than effects computed using the epistatic model (r=0.50).

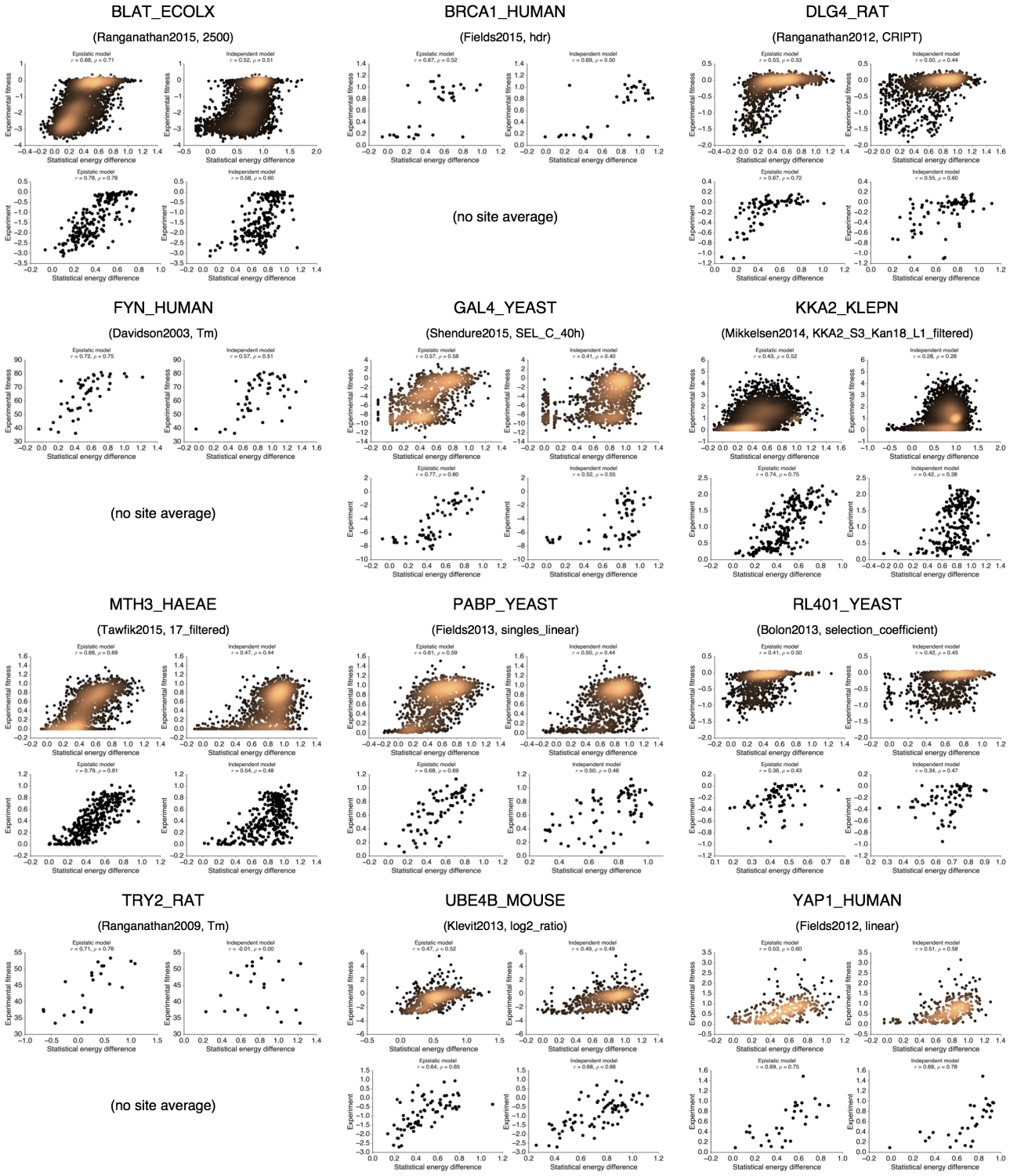

Figure A.2.: Agreement between computed and experimental mutation effects. Relationship between mutation effects computed using the epistatic and independent models (left and right plots) for individual mutants and average mutational sensitivities per site (top and bottom plots; orange corresponds to higher density of points; see Table A.1 for analyzed subsets of mutants). The plot for PABP_YEAST does not include a single outlier of high experimental fitness, which was still used in the correlation calculations.

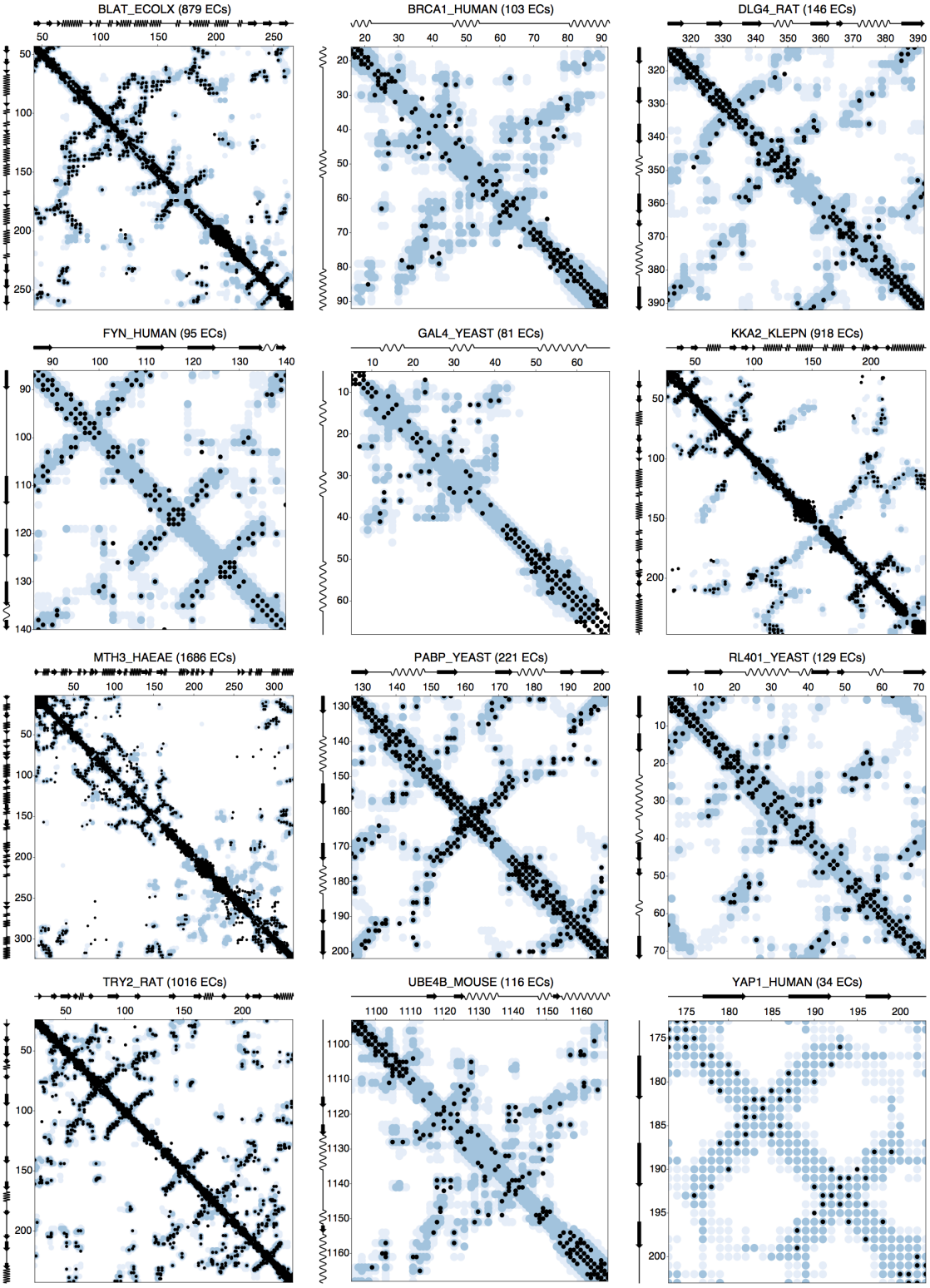

Figure A.3.: Evolutionary couplings correspond to residue contacts. Many evolutionarily coupled pairs (black dots: all significant pairs with quality score above background noise) are close in the 3D structure of the protein (dark and light blue: 5 and 8 Å minimum atom distance thresholds).

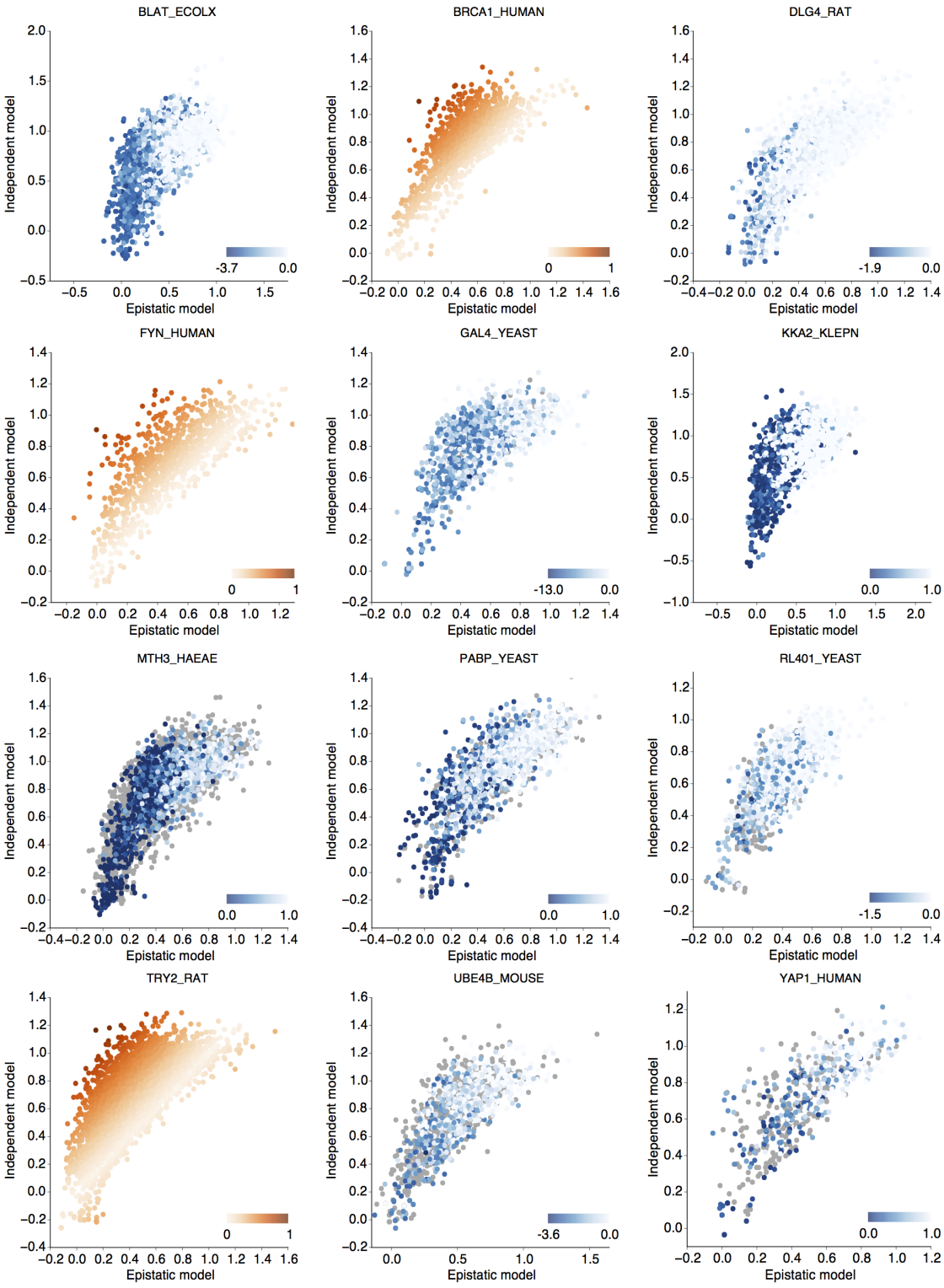

Figure A.4.: Mutation effects predicted differently between epistatic and independent models. The epistatic model tends to predict mutations as more damaging than the independent model, which wrongly quantifies experimentally deleterious mutations as neutral in many cases (white: no experimental effect, blue: deleterious, gray: no data available). For proteins with low-throughput mutagenesis data, scatters are colored in shades of orange according to the magnitude of difference between the models instead.

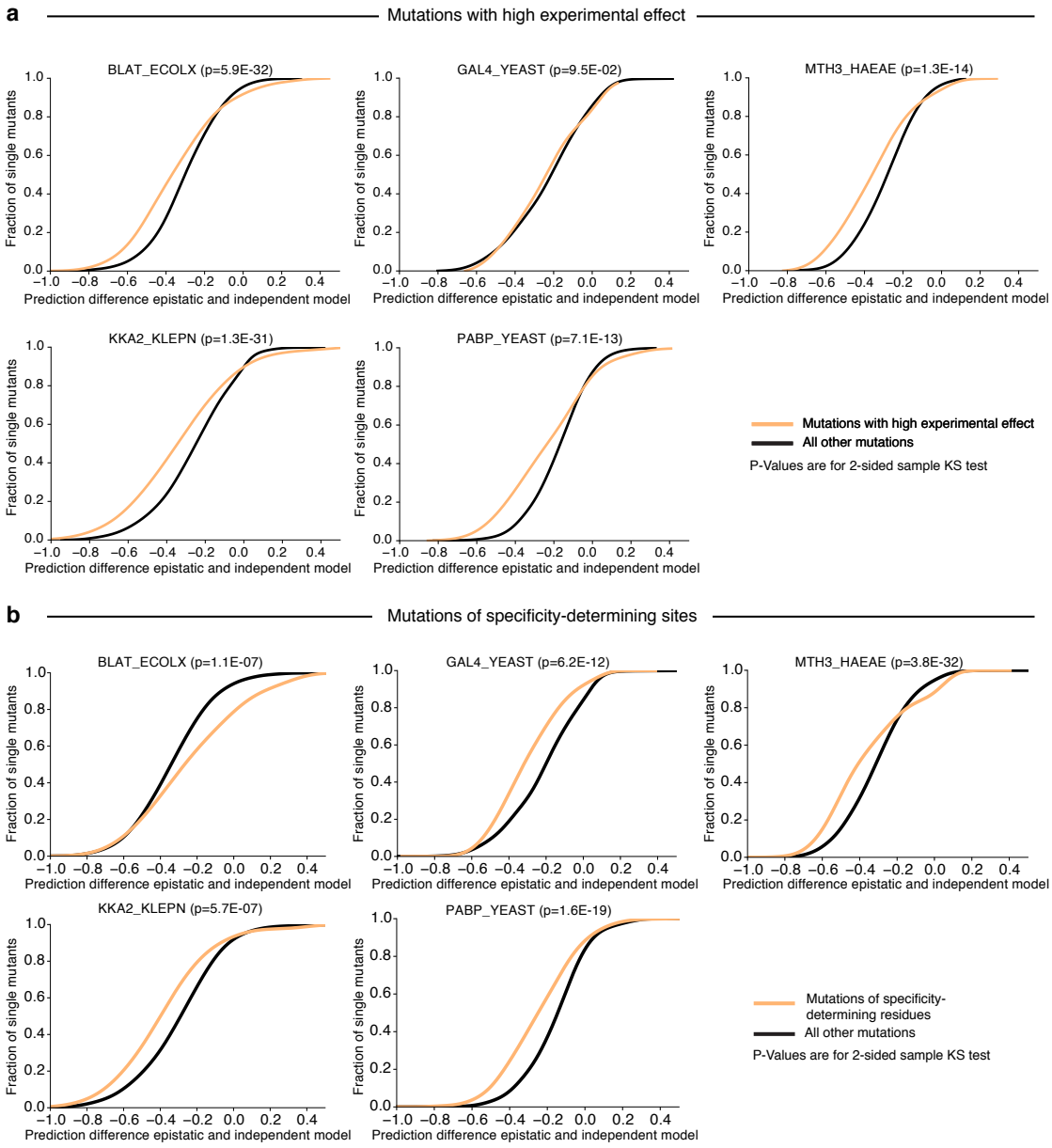

Figure A.5.: High-effect and specificity sites predicted more differently between models than others. Computed single mutation effects for (**a**) experimentally deleterious mutations (orange curve, determined by fitting a two-component Gaussian mixture model) and (**b**) mutations in specificity-determining sites (orange curve) are predicted more differently between the epistatic and independent models than all other remaining single mutations to the protein, respectively (black curves).

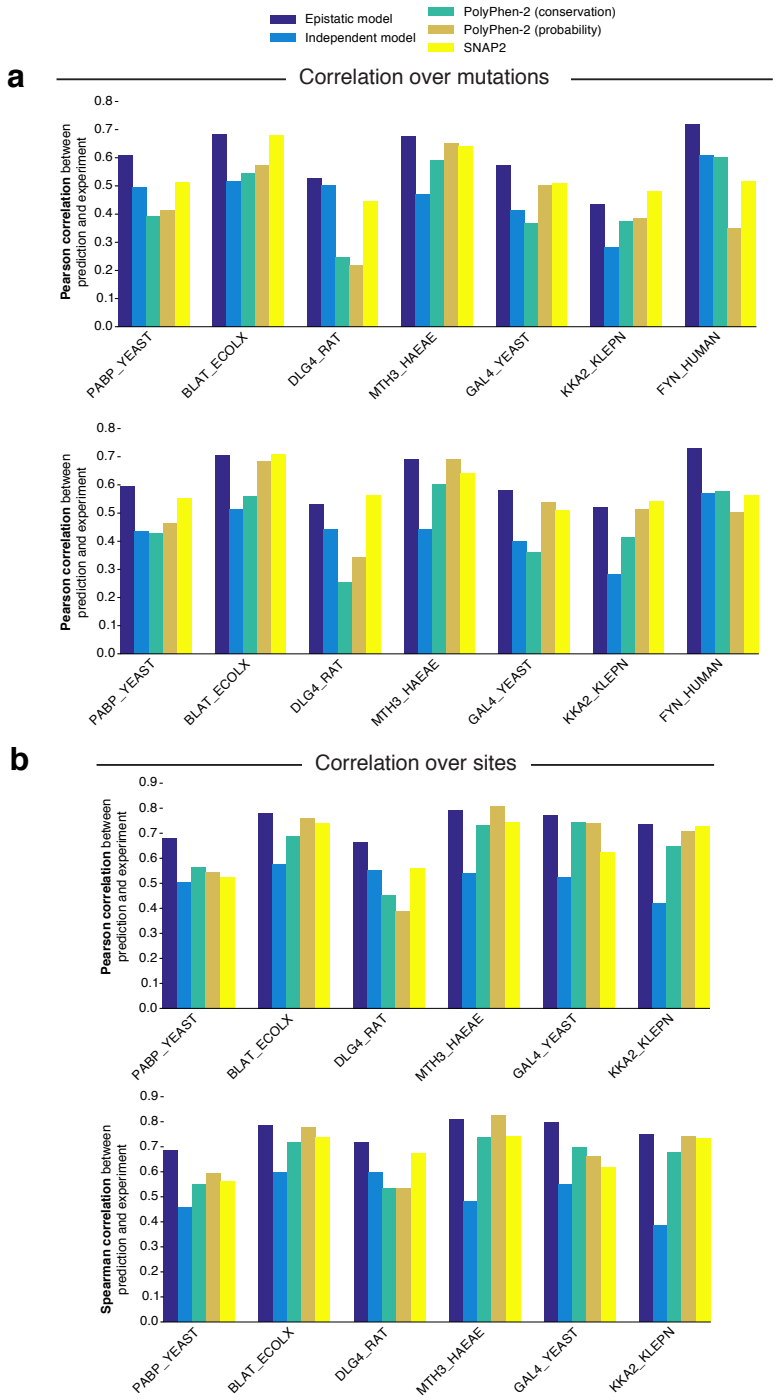

Figure A.6.: Comparison with machine learning-based prediction methods. Correlations between predicted and experimental mutation effects when comparing (**a**) individual mutations and (**b**) average mutational sensitivities per site (top panels: Pearson r, bottom panels: Spearman ρ) for a selection of state-of-the-art machine learning methods and our sequence-based approaches.

B. Publications

The following published first-author manuscripts have been appended to this thesis:

Hopf, T. A. *et al.* Sequence co-evolution gives 3D contacts and structures of protein complexes. *Elife* **3** (2014) [i]

Hopf, T. A. *et al.* Amino acid coevolution reveals three-dimensional structure and functional domains of insect odorant receptors. *Nat Commun* **6,** 6077 (2015) [ii]

Summaries and author contributions for each of the manuscripts are given on the subsequent pages.

[i] Reproduced from doi:10.7554/eLife.03430 under Creative Commons CC BY 4.0 license

[ii] Originally published in the journal *Nature Communications* and reproduced according to the author reuse guidelines of Nature Publishing Group

Sequence co-evolution gives 3D contacts and structures of protein complexes

Thomas A. Hopf*, Charlotta P.I. Schärfe*, João P.G.L.M. Rodrigues*, Anna G. Green, Oliver Kohlbacher, Chris Sander, Alexandre M.J.J. Bonvin & Debora S. Marks
* Joint first authors

Protein-protein interactions are molecular phenotypes relevant to most cellular processes. Yet, structural information is only available for a small subset of all known interactions, creating a need to close this gap using computational predictions. We developed a method that infers the molecular details of protein interactions from evolutionary sequence covariation making no use of solved structural templates. The approach identifies putatively interacting pairs of sequences per species using a strategy that is based on the genomic proximity of the interaction partners and then computes coevolving residue pairs between them (inter-protein evolutionary couplings). Assuming that the coevolutionary signal occurs due to spatial closeness of residue pairs, distance restraints between the interacting subunits can be used to assemble the complex from its constituents. To test the validity of our method, we compiled a comprehensive set of bacterial protein interactions of known structure where both subunits are proximal on the genome and predicted inter-protein couplings for all examples that had sufficient sequence information available. We found that residue pairs with significant coupling scores, as quantified by a newly developed score for quality assessment, frequently correspond to contacts in the 3D structures of the complexes and critical functional interactions. These couplings allowed to accurately reconstruct the complex 3D structures using standard biomolecular docking algorithms. Besides the identification of interacting residue pairs and structure prediction, we asked if our approach can be used to predict if two proteins interact or not. As a case study, we analyzed the bacterial ATP synthase complex and successfully classified most pairs of subunits as interacting or non-interacting. Motivated by the performance on solved examples, we then applied our method to the *de novo* prediction of protein complexes without known structure, and obtained biologically plausible evolutionary couplings and models for many of the candidates. Prominent examples include the elusive interaction between the subunits *a* and *b* of ATP synthase, and the stress/SOS response complex UmuCD. Despite the currently limited applicability of the method to complexes with enough paired sequences, we anticipate that it will provide access to study many protein interactions currently not solved by experiment.

TAH, CPIS, DSM: Conception and design, Acquisition of data, Analysis and interpretation of data, Drafting or revising the article; JPGLMR: Assisted docking in Haddock; AGG: Conducted comparison of ATP synthase subunit interactions; OK: Acquisition of data, Analysis and interpretation of data; CS: Conception and design, Drafting or revising the article; AMJJB: Provided expertise on Haddock docking protocols, Drafting or revising the article.

Amino acid coevolution reveals three-dimensional structure and functional domains of insect odorant receptors

Thomas A. Hopf, Satoshi Morinaga, Sayoko Ihara, Kazushige Touhara, Debora S. Marks & Richard Benton

Insect olfactory receptors are a large family of α-helical transmembrane proteins responsible for the molecular recognition of odors and subsequent triggering of a neuronal response. Despite great interest in identifying the exact signaling mechanism of these ion channel proteins, the molecular details of signal transmission remain unknown. Three-dimensional structure information would provide an opportunity to obtain a more detailed understanding of olfactory receptors, but is not available for any member of the entire protein family. To provide structural information, we extended our previously developed method for predicting three-dimensional structures of transmembrane proteins from evolutionary sequence covariation. The approach identifies coevolving positions (evolutionary couplings) in protein family alignments using a pairwise undirected graphical model of sequences and assumes that amino acids covary due to spatial proximity of the residues. By defining distance restraints on strongly coupled pairs of positions, three-dimensional structure models can be obtained using standard distance geometry and simulated annealing algorithms. We then calculated evolutionary couplings and three-dimensional models for two members of the olfactory receptor family, the co-receptor ORCO and the ligand-specific receptor OR85b. The predicted structures show a novel three-dimensional fold with a packing arrangement of the seven transmembrane helices that is distinct from other known membrane protein folds. Importantly, the olfactory receptors have a different fold than G-protein coupled receptors, which the insect odor receptors were presumed to belong to. To verify the validity of the model and gain insights in the spatial organization of olfactory receptor function, we calculated positions that are particularly strongly coupled to other positions, and compared them with known functional residues from mutational studies. We found that strongly coupled and known functional residues cluster in 3D in the same parts of the molecule, suggesting that the models are plausible and these are distinct functional domains of the receptors. We additionally tested the role of the strongly coupled N-terminal region experimentally and found that it plays an essential role for receptor function. Our results provide a first detailed insight in olfactory receptor structures and a model for further studies of this important class of proteins.

T.A.H. developed analysis methods, performed multiple sequence alignments, EC calculations and model building, and contributed to the writing of the manuscript. D.S.M. developed analysis methods, analyzed data and contributed to the writing of the manuscript. S.M. designed and performed the OR experimental analysis, and S.I. and K.T contributed to experimental design and interpretation. R.B. conceived the project, annotated gene sequences and models, analyzed data and wrote the manuscript.