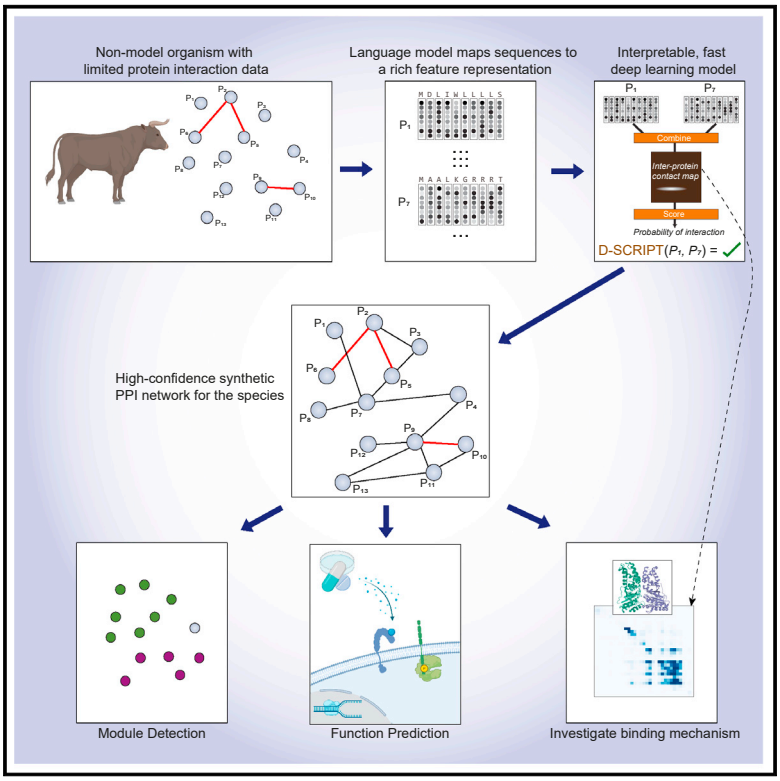
Article

Cell Systems

D-SCRIPT translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein-protein interactions

Graphical abstract



Highlights

- Method to predict protein-protein interactions from primary amino acid sequences
- Resulting predictions enable network clustering and functional module detection
- Efficient genome-scale PPI prediction helps to tackle the genome-to-phenome problem
- Application in bovine rumen reveals links between metabolism and the immune system

Authors

Samuel Sledzieski, Rohit Singh, Lenore Cowen, Bonnie Berger

Correspondence

cowen@cs.tufts.edu (L.C.), bab@mit.edu (B.B.)

In brief

With the explosive growth of available gene sequences, the ability to predict the interactome of an organism from its genome would help address the pressing challenge of genome-to-phenome mapping. Our method, D-SCRIPT, leverages advances in deep language models to map protein sequences to implicit structure representations and predict interaction between two proteins based on their structural compatibility. D-SCRIPT's structure-aware approach generalizes to unseen species better than current approaches, and its efficiency allows for rapid genome-scale investigation of protein function.

Sledzieski et al., 2021, Cell Systems *12*, 969–982
October 20, 2021
https://doi.org/10.1016/j.cels.2021.08.010

CellPress

Cell Systems
Focus on RECOMB

CellPress
OPEN ACCESS

Article

D-SCRIPT translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein-protein interactions

Samuel Sledzieski,[1,4] Rohit Singh,[1,4] Lenore Cowen,[2,*] and Bonnie Berger[1,3,5,*]
[1]Computer Science and Artificial Intelligence Lab., Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[2]Department of Computer Science, Tufts University, Medford, MA 02155, USA
[3]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[4]These authors contributed equally
[5]Lead contact
*Correspondence: cowen@cs.tufts.edu (L.C.), bab@mit.edu (B.B.)
https://doi.org/10.1016/j.cels.2021.08.010

SUMMARY

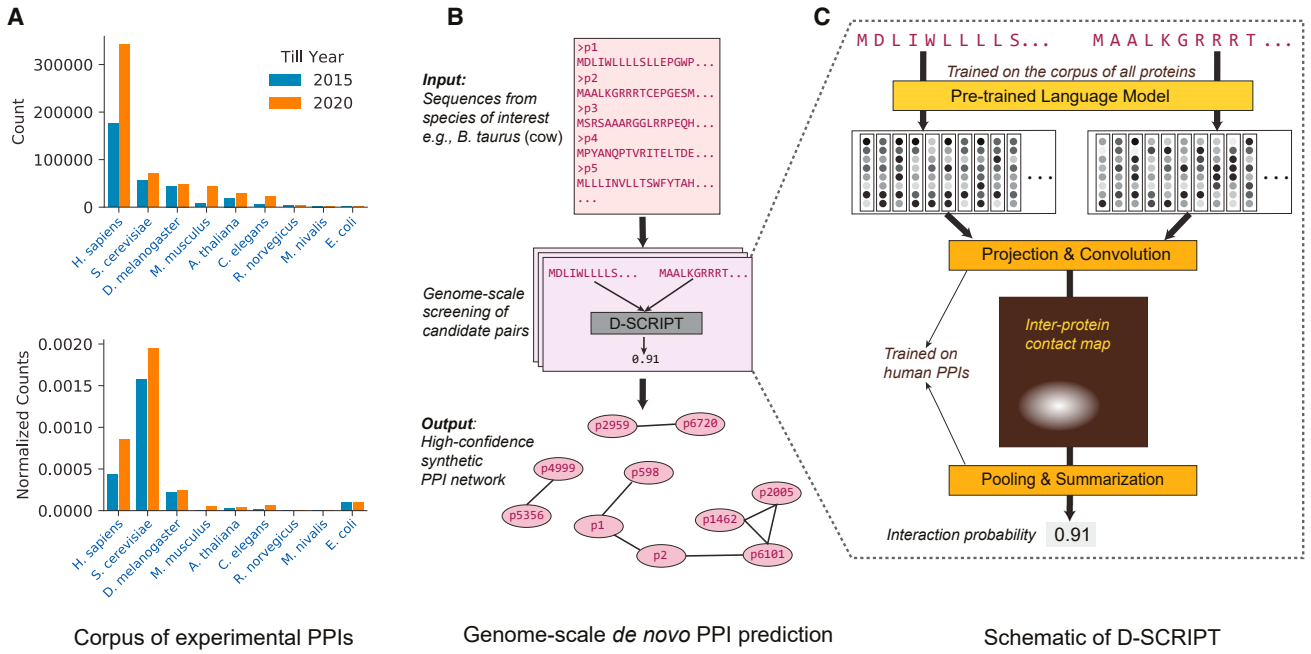
We combine advances in neural language modeling and structurally motivated design to develop D-SCRIPT, an interpretable and generalizable deep-learning model, which predicts interaction between two proteins using only their sequence and maintains high accuracy with limited training data and across species. We show that a D-SCRIPT model trained on 38,345 human PPIs enables significantly improved functional characterization of fly proteins compared with the state-of-the-art approach. Evaluating the same D-SCRIPT model on protein complexes with known 3D structure, we find that the inter-protein contact map output by D-SCRIPT has significant overlap with the ground truth. We apply D-SCRIPT to screen for PPIs in cow (*Bos taurus*) at a genome-wide scale and focusing on rumen physiology, identify functional gene modules related to metabolism and immune response. The predicted interactions can then be leveraged for function prediction at scale, addressing the genome-to-phenome challenge, especially in species where little data are available.

INTRODUCTION

The systematic mapping of physical protein-protein interactions (PPIs) in the cell has proven extremely valuable in deepening our understanding of protein function and biology. In species such as yeast and human, where a large network of experimentally determined PPIs exists (Fields and Song, 1989; Kumar and Snyder, 2002; Krogan et al., 2006; Taipale et al., 2014; Sahni et al., 2015), this PPI network information has proven valuable for downstream inference tasks in understanding functional genomics and biological pathway analysis (Sharan et al., 2007; Navlakha and Kingsford, 2010; Cho et al., 2016; Cowen et al., 2017; Choobdar et al., 2019). However, in most species—especially, non-model organisms—the coverage of experimental PPI data remains very low (Figure 1A).

Computational prediction of PPIs can help mitigate the lack of experimental data and facilitate biological discovery. Although substantial progress has been made in PPI prediction overall, the notable case of *de novo* prediction for less-studied proteins and non-model organisms continues to be a challenge. The lack of functional genomic data in such situations makes it difficult to apply methods based on bootstrapping from the connectivity patterns of known PPIs (Hosur et al., 2012; Lei and Ruan, 2013; Hulovatyy et al., 2014; Kovács et al., 2019; Devkota et al., 2020) or those that infer PPIs from other protein-protein association modalities such as co-expression and co-localization (Cho et al., 2016; Franz et al., 2018; Wang et al., 2018; Szklarczyk et al., 2019). Recently, deep-learning-based methods have offered the prospect of predicting PPIs just from sequence data. Unfortunately, existing models (Hashemifar et al., 2018; Chen et al., 2019) have shown limited generalizability: they work quite well when applied to the species they were trained on, but their performance declines in a cross-species context.

Here, we introduce D-SCRIPT (deep sequence contact residue interaction prediction transfer, see Box 1), a structure-aware deep-learning approach to PPI prediction with a geometrically interpretable neural network architecture that is able to make meaningful PPI predictions in the cross-species setting. Our key conceptual advance is implementing an interpretable, structure-based model despite only having sequence-based inputs: a well-matched combination of input featurization and neural network architecture allow for D-SCRIPT to be trained solely from sequence data, supervised only with a binary interaction label and yet produce an intermediate representation that substantially captures the structural mechanism of interaction between the protein pair. Leveraging recent advances in protein language modeling, we first apply Bepler and Berger's



Corpus of experimental PPIs

Genome-scale *de novo* PPI prediction

Schematic of D-SCRIPT

Figure 1. D-SCRIPT motivation and workflow

We demonstrate how D-SCRIPT can be used genome-wide to predict a complete PPI network in the fly.

(A) Experimentally derived PPI data are scarce in species outside of human and yeast, even when normalized for size of the genome (sourced from BioGRID, STAR Methods).

(B) A D-SCRIPT model, after being trained on a large corpus of human PPI data, can be broadly applied to a species of interest even if little PPI data are available in that species. For each pair of proteins in the target species, D-SCRIPT converts the pair of protein sequences into a score representing probability of interaction. Because D-SCRIPT scales to large numbers of protein pairs and maintains performance across species, it can be used to score all protein pairs genome-wide to predict a synthetic PPI network in the species, facilitating a genome-to-phenome translation.

(C) Blowup detail of the D-SCRIPT architecture from the box in (B) (Figure 2 for more detail). D-SCRIPT generalizes due to its structurally motivated design. The pre-trained language model generates structural features for a single protein, whereas the projection and convolution model the interaction between every pair of residues in the candidate pair. In the final layer, we introduce a magnitude regularization term to ensure the prediction of an inter-protein contact map that is structurally plausible.

deep-learning-based language model of single proteins to construct our features (Bepler and Berger, 2019). Using this pre-trained model results in informative protein embeddings (i.e., representations in a high-dimensional space) that are implicitly endowed with structural information about each of the proteins. D-SCRIPT's generalizability and interpretability then come from its ability to learn informative geometric representations of the proteins, individually and jointly. In particular, it learns how to transform the two protein embeddings into a 2D contact map, encoding the intuition that a physical interaction between two proteins requires that a subset of the residues in each protein be in contact with the other protein (Figures 1B and 1C).

Evaluating D-SCRIPT in the cross-species prediction setting, where a method trained on human PPIs is used to predict PPIs in several less-studied model organisms, we show that it substantially improves upon existing methods, including the state-of-the-art deep-learning method PIPR (Chen et al., 2019) in a stringent cross-validation experiment. In addition to comparing the accuracy of PPI predictions in cross validation, we demonstrate that the interpretability and downstream utility of D-SCRIPT results in several ways. First, we demonstrate that on a genome-wide scale, *de novo* PPIs predicted by D-SCRIPT produce a network whose modular structure produces clusters of proteins with greater functional coherence than those produced from PIPR predictions. Next, on assessing the physical plausibility of the intermediate contact map representation, we find that the map partially discovers the structural mechanism of an interaction despite the model having been trained only on sequence data. Specifically, we evaluate our predictions on Hwang et al.'s benchmark database of 3D structures of docked protein pairs and observe that our model's predicted contact map is substantially similar to the ground-truth inter-protein contact map in cases where our model predicts an interaction (Hwang et al., 2010).

To demonstrate the utility of D-SCRIPT as a tool to study novel systems in less-studied organisms, we investigate the rumen in *Bos taurus* (cow). We apply D-SCRIPT to predict new PPIs across a large subset of bovine proteins and decompose the network of D-SCRIPT predicted PPIs into functional gene modules. Starting from a seed set of genes found by Jiang et al. (Jiang et al., 2014) to be overexpressed in the rumen, we identify five functional gene modules involved in cellular metabolism and growth, immune response, and transcriptional regulation, suggesting links between metabolism and transcriptional regulation through *MRPL4* and *H15* domain-containing proteins.

Box 1. Progress and potential

Progress: Almost all cellular processes involve proteins interacting with each other in three-dimensional space; protein-protein interactions translate genomic information into biological function. Although availability of direct, large-scale measurements of protein-protein interactions (PPIs) islimited, genome sequences are available at unprecedented scale. Thus, there is a need for scalable methods that predict interactions from protein sequences, facilitating their functional characterization.

The key conceptual advance of D-SCRIPT is accurately modeling three-dimensional, physical intuition about protein structure and function with just one-dimensional, sequence-based inputs. To do so, we leverage a deep protein language model that maps protein sequences to a high-dimensional representation that automatically captures structural features of proteins. We combine that deep protein language model with a carefully crafted neural network that capitalizes on these features to predict protein-protein interactions. We achieve D-SCRIPT's generalizability by basing our prediction on structural compatibility, with the intuition that the structural bases of protein interaction are similar across species: when two proteins interact, their structures bind in a thermodynamically favorable way. Accordingly, D-SCRIPT substantially outperforms state-of-the-art approaches when applied to out-of-sample species (e.g., predicting PPIs in fly after being trained on human PPI data).

Our model is not a black box—for every prediction, D-SCRIPT also outputs an inter-protein contact map, identifying the likely binding residues of the two proteins. Our approach is inspired by the remarkable success of adapting pre-trained language models via transfer learning in the domains of speech recognition, translation, and natural language processing.

Potential: With advances in sequencing technologies, genomes for many organisms are now available (RefSeq currently lists 5,963 eukaryotic genome assemblies). However, for most non-model organisms, little is known about the functions of specific genes, with functional genomic data being rarely available. D-SCRIPT offers one way to address the "genome-to-phenome" challenge in these species. Because D-SCRIPT generalizes to new out-of-sample species, it can be applied out of the box to predict PPIs *de novo* from the genome of a newly sequenced organism, where it is fast enough that a genome-wide PPI prediction screen can be performed in a few days. Standard graph-theoretic analysis of the predicted PPI network, weighted by confidence in interaction, can then be applied to identify functional modules and annotate gene function at scale. We demonstrate this workflow through a study of protein function in cow (Bos taurus), with a focus on the proteins active in the cow rumen. We identified functional gene modules involved in cellular metabolism and growth, immune response, and transcriptional regulation.

RESULTS

Overview of the D-SCRIPT model

Our deep-learning model for predicting PPIs directly from protein sequences, similar to previous deep-learning methods DPPI (Hashemifar et al., 2018) and PIPR (Chen et al., 2019), is composed of two stages (Figure 1C). The first stage generates a rich feature representation for each protein separately, and the next stage estimates an interaction probability based on these features, with the model being trained end-to-end across both stages. In both DPPI and PIPR, much of the model complexity lies in the feature generation, which is learned *ab initio* from the training data.

D-SCRIPT differs from these approaches in the design and relative complexity of the two stages. First, we apply a pre-trained model to generate rich, structurally informative feature representations of the proteins (Figure 2). The pre-trained model was developed by Bepler and Berger, who built upon advances in deep-learning-based modeling of natural languages to design a language model for protein sequences : an n-amino acid protein sequence is mapped to an $n \times 6,165$ representation, with the various dimensions capturing local and global aspects of the protein structure (Bepler and Berger, 2019). We then learn a lower-dimensional projection ($n \times 100$) of this embedding as a compact representation for downstream interaction and structural prediction tasks. The second stage of D-SCRIPT encodes a structure-based model of protein interaction: in the contact module, the low-dimensional embeddings are used to compute an inter-protein contact map that corresponds to the locations of residue contacts between protein structures, and in the interaction module, this contact map is summarized into a single score (i.e., the probability of interaction). In each layer, the mathematical operations performed are rooted in structural intuitions. For example, to formalize the intuition that true-positive contact maps should be sparse but have isolated regions of strong contacts, we introduce a customized max-pooling operation and a magnitude regularization term in the loss. A more detailed description of our model architecture and training process is provided in STAR Methods.

D-SCRIPT generalizes well across species

We first sought to see how D-SCRIPT performed on the task of cross-species interaction prediction. We trained a model on human PPIs and evaluated it using PPI datasets from five other model organisms (STAR Methods). We compared D-SCRIPT with PIPR, shown by Chen et al. to be currently the best-performing sequence-based PPI prediction method (Chen et al., 2019), training both models on the same set of human PPIs; we compare their model complexity in the STAR Methods. In Table 1, we report the precision, recall, area under precision-recall curve (AUPR), and area under ROC curve (AUROC) of each method in each of five species. For highly unbalanced data, as is the case here, we note that AUPR is generally considered a better metric than AUROC. D-SCRIPT outperforms PIPR in a cross-species setting and maintains a high AUPR across all species, even those that are extremely evolutionary distant from human. In fact, its AUPR in these species remains comparable with that seen in human cross validation. Additionally, we compared with a hybrid approach (D-HYBRID) where PIPR was used to augment D-SCRIPT's prediction: when PIPR is highly confident that an interaction *does* occur ($\hat{p}>0.9$), the predicted probability of D-SCRIPT is increased by 50%. D-HYBRID

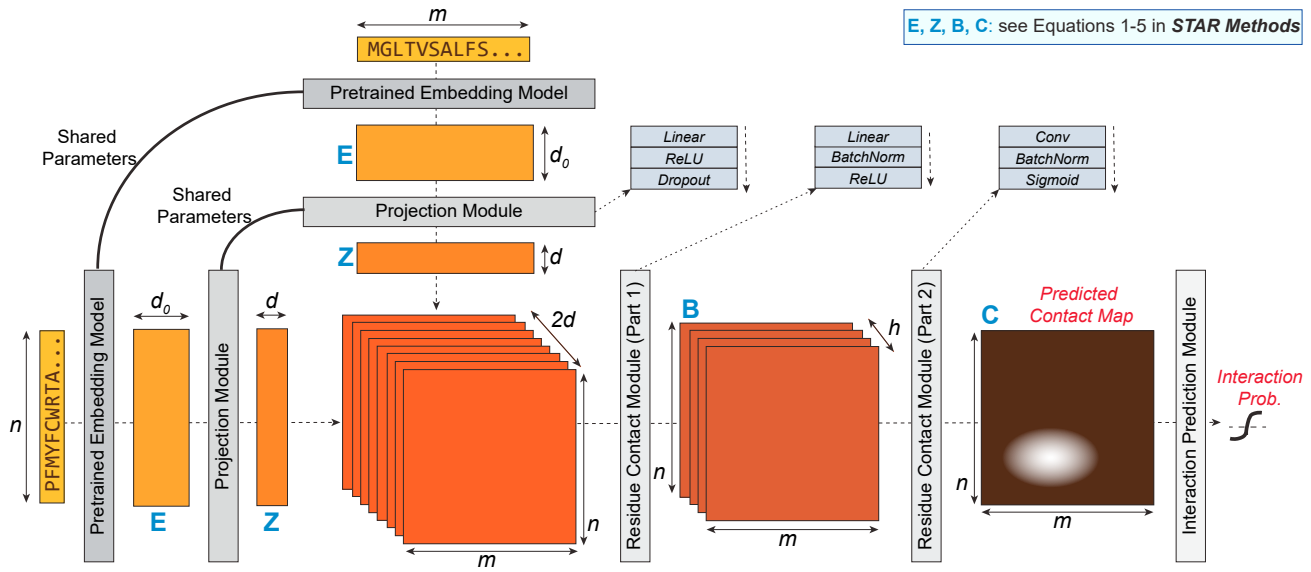


Figure 2. D-SCRIPT architecture

Left to right: the pre-trained embedding model, a deep-learning language model from Bepler and Berger, generates features for each individual protein. The projection module reduces them to *d* dimensions. Each low-dimensional single-protein embedding implicitly includes, among other features, an encoding that broadly captures the protein's residue-contact map (Figure 5). The contact module combines these low-dimensional embeddings to compute a sparse "inter-protein" contact map through a two-step process, which first computes a representation for each pair of residues, then incorporates local information using a convolutional filter. Finally, the interaction prediction module uses a customized max-pooling operation to predict the probability of interaction between the input proteins.

outperforms both D-SCRIPT and PIPR alone but improves D-SCRIPT only modestly in cross-species analysis. We also compare with a recently released method by Richoux et al. (2019) in Table S1.

Although our focus is on enhancing cross-species PPI prediction quality, we also sought to investigate how D-SCRIPT would perform at predicting within-species interactions in human. We performed 5-fold cross validation and report here the average across all folds. Table 1 shows that PIPR outperforms D-SCRIPT on human PPIs in cross validation. However, a hybrid approach works better in this case as well. Due to greater PIPR accuracy on in-sample species, we adapted the hybrid approach to use D-SCRIPT to augment PIPR's predictions (P-HYBRID): when D-SCRIPT is highly confident that an interaction does not happen ($\hat{p}$<0.01), we reduce the predicted probability from PIPR by half. Notably, P-HYBRID achieves substantially higher precision, although at the expense of recall. This may be a desirable trade-off in certain contexts, e.g., when generating PPI candidates for experimental validation. We note that, although the incremental performance of the hybrid models is modest, the observation that D-SCRIPT performs better on out-of-sample species, whereas PIPR performs better on in-sample species makes possible a simple combination of the two that does not substantially increase computation time and yet results in more accurate predictions both across and within species.

We further investigated the performance of D-SCRIPT and PIPR on subsets of the human data, seeking to better understand their relative strengths. Upon further analysis, we observe that D-SCRIPT performs better on interactions involving proteins that occur infrequently in the PPI network, whereas PIPR performs better on those involving proteins that occur frequently (STAR Methods; Table S2). This suggests that PIPR may perform better when a large amount of training data are available, but D-SCRIPT may generalize better to new proteins (and species) (Figure S1).

D-SCRIPT predictions are functionally informative

The importance of PPI networks arises, in part, from the graph-theoretic analyses on them, which enable the functional characterization of un-annotated proteins. Therefore, we sought to test if D-SCRIPT's success at cross-species generalization would translate to better functional inference in new species. In particular, we hypothesized that, compared with PIPR, the D-SCRIPT model trained on human data should facilitate more accurate inference of protein functional modules in *Drosophila melanogaster*. Toward this, we generated a set of 10,475,595 candidate pairs from the set of *D. melanogaster* proteins in STRING. Using D-SCRIPT and PIPR's human-trained models, we predicted interactions over this candidate set. On the resulting PPI networks, we performed functional module detection and quantified the functional coherence of 374 (PIPR) and 384 (D-SCRIPT) modules using available GO (gene ontology) annotations from FlyBase (Thurmond et al., 2019) (STAR Methods). Functional coherence of a module quantifies the extent to which proteins in the module are likely to participate in the same biological functions. A higher average within-cluster similarity is desirable because it enables more accurate functional characterization of novel proteins by associativity and discovery of protein functional modules. We find that the average within-cluster similarity

Table 1. Evaluation of models trained on human PPIs

Species	Model	AUPR	Precision	Recall	AUROC
M. musculus	PIPR	0.526	0.734	0.331	**0.839**
	D-SCRIPT	0.580	0.818	0.346	0.833
	D-HYBRID	**0.609**	**0.820**	**0.355**	0.838
D. melanogaster	PIPR	0.278	0.521	0.121	0.728
	D-SCRIPT	0.552	**0.798**	0.359	**0.824**
	D-HYBRID	**0.562**	**0.798**	**0.361**	**0.824**
C. elegans	PIPR	0.346	0.673	0.142	0.757
	D-SCRIPT	0.548	0.840	0.306	0.813
	D-HYBRID	**0.559**	**0.841**	**0.308**	**0.814**
S. cerevisiae	PIPR	0.230	0.398	0.085	0.718
	D-SCRIPT	0.405	0.706	0.223	**0.789**
	D-HYBRID	**0.417**	**0.708**	**0.225**	**0.789**
E. coli	PIPR	0.308	0.629	0.131	0.675
	D-SCRIPT	0.571	0.791	**0.520**	**0.863**
	D-HYBRID	**0.588**	**0.793**	0.394	**0.863**
H. sapiens (5-fold cross validation)	PIPR	0.835	0.838	**0.701**	0.960
	D-SCRIPT	0.516	0.728	0.278	0.833
	P-HYBRID	**0.844**	**0.949**	0.400	**0.962**

We show performance of D-SCRIPT, PIPR (currently the best-performing sequence-based deep-learning PPI prediction method), and two hybrid approaches: D-HYBRID refers to D-SCRIPT predictions augmented with PIPR predictions, and P-HYBRID refers to PIPR predictions augmented with D-SCRIPT predictions (STAR Methods). *H. sapiens* results are average performance over 5-fold cross validation, using 38,345 positive (and ten times as many negative) PPIs for training for each fold. All other species were evaluated using a model trained on human data and evaluated on 5,000 positive and 50,000 negative PPIs (2,000/20,000 for *E. coli* due to limited data). D-SCRIPT outperforms PIPR cross-species, although PIPR performs better on in-sample species (i.e., human cross validation). The best-performing method for each species and metric is bolded. AUPR, area under precision-recall curve; AUROC, area under receiver operating characteristic curve. See also Table S1.

when interactions are predicted using D-SCRIPT is significantly higher than when using PIPR ($p = 0.000723$, one-tailed t test). Compared with PIPR, D-SCRIPT also results in 24% more highly coherent clusters (Jaccard similarity ≥ 0.468, 90th percentile) (Figures 3 and S2). In addition to this cluster-based test, we directly compared the functional similarities between protein pairs (measured as the overlap of GO functional annotations) with their graph-theoretic similarities implied by the D-SCRIPT and PIPR networks (STAR Methods). We find that D-SCRIPT admits a significantly stronger correlation between the two measures than PIPR (Spearman ρ, 0.123 versus 0.005, $p = 0.00$, Fisher r-to-z).

We additionally applied D-SCRIPT and PIPR to computationally screen all SARS-Cov-2 proteins against 19,777 human proteins, predicting approximately 3,000 viral-host PPIs from each method and characterized each viral protein's function by the GO annotations of its human interactors (STAR Methods). We find that compared with the corresponding annotations derived from 332 experimentally determined PPIs (Gordon et al., 2020), D-SCRIPT-based annotations overlap more with the experimental results than those from PIPR ($p = 0.059$, paired one-tailed t test).

Case study: Protein function and interaction in the bovine rumen

Because D-SCRIPT generalizes well to species with limited available PPI data, it enables the study of protein functional pathways through *de novo* prediction of protein interaction networks. Following this, we undertook a study of protein interaction in the bovine rumen to investigate the biological processes involved in rumination. In a comprehensive study of the sheep (*Ovis aries*) genome, Jiang et al. identify several genes that are preferentially expressed in rumen tissue, including *PRD-SPRRII*, *S100-A2*, *S100-A12*, and *TCHHL2* (Jiang et al., 2014). Using BLAST (Altschul et al., 1990), we identified 12 putative homologous proteins in the ARS-UCD1.2 cow (*Bos taurus*) genome assembly (Table S3) to focus on in our analysis. Including these rumen-specific homologs, we selected 24,195 bovine proteins and used the human-trained D-SCRIPT model to predict the probability of interaction for fifty million candidate pairs. We predicted a network of 476,399 positive interactions between 17,811 proteins and performed functional module detection and gene set enrichment analysis (STAR Methods). To quantitatively assess the accuracy of our predicted edges and clusters, we computed the co-expression of each pair of genes (Figure 4F) in 93 tissue samples. We find that pairs of genes for which we predict an edge with D-SCRIPT are significantly more likely to be coexpressed than a random pair of genes ($p <$ 1e-84, one-sided t test). Further, the correlation between expression vectors is even stronger for pairs that appear in the same functional module, even in cases where D-SCRIPT does not predict an interaction between the proteins.

The largest cluster we identify (cluster A, Figure 4A) comprises 65 proteins, including two homologs of *PRD-SPRRII*. 34 of the genes in cluster A are homologous to various human protein

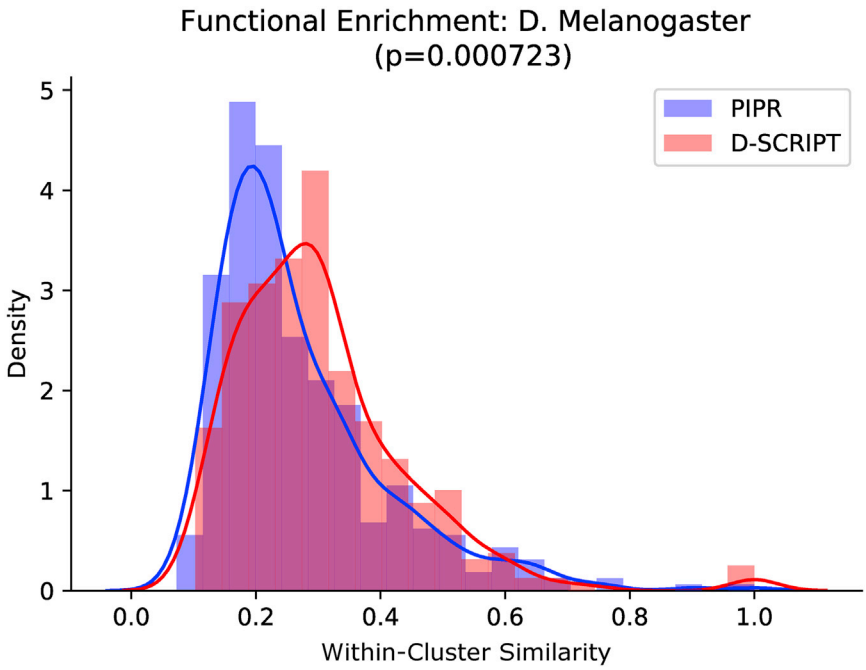


Figure 3. Improved protein functional characterization using D-SCRIPT modules

D-SCRIPT recovers more functionally coherent clusters than PIPR ($p = 0.000723$, one-tailed t test). 384 (374) protein clusters were generated by evaluating 10,475,595 candidate protein pairs with D-SCRIPT (PIPR). We computed the diffusion-state distance (DSD) between all proteins, clustered the DSD matrix using spectral clustering, filtered out small (< 3) clusters, and recursively split large (> 100) clusters. Within-cluster similarity was calculated as the average Jaccard similarity between GO Slim annotations of all pairs of proteins in the cluster. See also Figure S2.

tyrosine phosphatases, across multiple classes of protein tyrosine phosphatases categorized in earlier work (Alonso et al., 2004; Alonso and Pulido, 2016). Other genes in the cluster are *TXNL1*, known to buffer response to oxidative stress (Yu et al., 2019; Zhao and Qi, 2021), and serine/threonine protein kinases, including *VRK3* (Lee et al., 2017) and *STK33* (Brauksiepe et al., 2008). Two genes in cluster A have previously been associated with disease in cattle, *TUBD1* (Schwarzenbacher et al., 2016), where missense mutations were associated with increase juvenile mortality in cattle, and *NUAK1* (Sasaki et al., 2014), which was found to be differentially expressed in cows with milk fever.

This cluster is connected through *PTPMT1* and *TARS* to a cluster of 53 proteins (cluster B, Figure 4B), which are active in metabolism within the mitochondria, and participate in oxidoreductase and transaminase catalytic activity. This cluster also contains one homolog of *PRD-SPRRII*. Ruminants, such as the cow, exhibit patterns of energy metabolism that are quite different from non-ruminants such as humans or rodents (Dodson et al., 2010). Because ingested carbohydrates are fermented to short chain fatty acids in the rumen, glucose demand is met by gluconeogenesis, controlled by transcriptional regulation, and the associated genes and pathways are implicated in metabolic disorders that affect dairy cows such as fatty liver and ketosis (Aschenbach et al., 2010). Many of the human homologs of genes in cluster B are known to localize to the mitochondria, including *ACSS1* (Castro et al., 2012), *AGXT2* (Rodionov et al., 2014), *COA6* (Soma et al., 2019), *DECR1* (Kamiński *et al.*, 2009), *MTHFD2* (Zhu and Leung, 2020), *OAT* (Ginguay et al., 2017), plus *LHPP*, which was predicted by Gohla (2019) to be involved in mitochondrial oxidative phosphorylation. Other genes in the cluster have been implicated in mitochondrial tRNA modification, including *CDK5RAP1* (Reiter et al., 2012), *MTO1* (Ghezzi et al., 2012), *TARS* (Chen et al., 2018), and *TFAM* (Pohjoismäki et al., 2006). Still others, such as *ABAT* (Besse et al., 2015), *FOXRED1* (Fassone et al., 2010), and *NDUFV1* (Srivastava et al., 2018), have human homologs whose role in mitochondrial rare diseases has been documented. Several of the genes in this cluster have been implicated as important trait or disease markers in dairy cows, pigs, and sheep, making the predicted interactors of these genes of particular interest. For example, *BDH1*, involved in ketogenesis (Gao and Oba, 2016) has been linked by multiple studies to health of lactating dairy cows (Zarrin et al., 2014; Gao and Oba, 2016). High expression of *BDH1* was positively correlated with milk yield and negatively correlated with fat yield in buffalo (Yadav et al., 2015). Polymorphisms of the *DECR1* gene have been connected to meat quality (Jing et al., 2009; Kamiński et al., 2009), and *AHCYL2* was one of ten candidate disease genes suggested to be involved in susceptibility to DA (Huang et al., 2019).

This cluster is further connected through *MRPL4* and two *H15* domain-containing proteins to a module of 55 proteins (cluster E, Figure 4E) involved in transcriptional regulation, with significant enrichment for nucleosome and chromatin assembly and organization. In Wei et al. (Wei et al., 2013), *MRPL4* was identified as being involved in immune and inflammatory pathways, where the human homolog was suggested as a disease gene for allergic rhinitis. Of the set of highly expressed sheep rumen protein homologs, this cluster contains all four homologs of *S100-A12* and one of *S100A-2*. It also contains homologs to other human S100 proteins, *S100-A4*, *S100-A7*, *S100-A8*, *S100-G*, *S100-A11*, and *S100-A16*. Many of these S100 proteins have been implicated in progression of human epithelial tumor progression, cell differentiation, and chronic inflammation (Heizmann, 2019). A sub-cluster of three of these proteins, S100-A12, S100-A7, and S100-A8, has been implicated in the innate immune response to pathogens, including parasites *E. coli* and *H. pylori* (Hsu et al., 2009), where some have been shown to function in the nutritional immunity mechanism by outcompeting bacterial metal ion transporters (Kozlyuk et al., 2019). However, for the S100 family of proteins, there should be some caution in assuming specificity of function translates across species: for example, the bovine S100-G most likely buffers calcium but is not likely to be a calcium sensor such as

CellPress
OPEN ACCESS

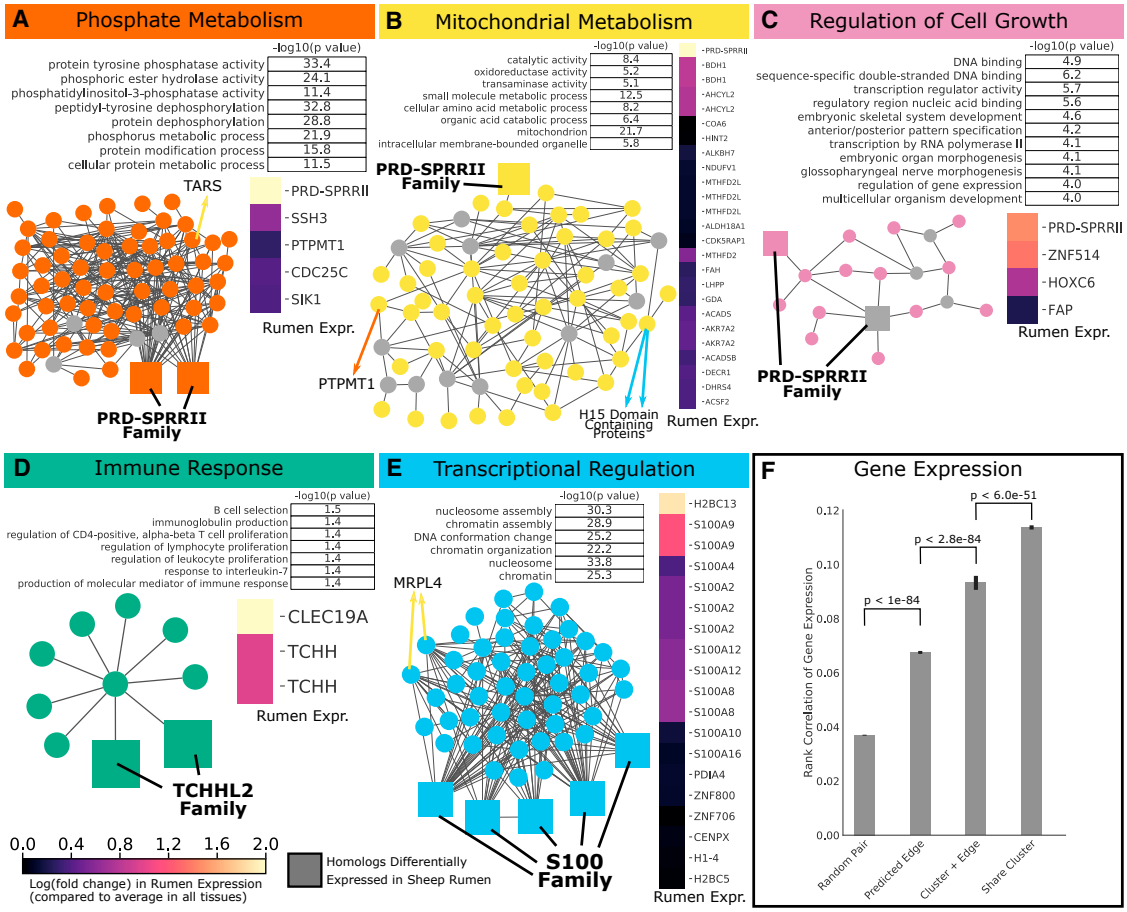


Figure 4. Protein interaction network in bovine rumen

We applied D-SCRIPT to predict a *de novo* PPI network in cow (*B. taurus*) and investigated specifically the functional modules likely to be active in the cow rumen (A–D). After evaluating 50 million candidate protein pairs, we generated a network of 476,399 predicted PPIs between 17,811 proteins and performed spectral clustering on the diffusion-state distance (DSD) matrix of the network to identify functional modules, shortlisting five modules related to rumen physiology. A recent RNA-seq study validates several proteins in these modules as being strongly overexpressed in rumen tissue. For each module, we report gene ontology molecular function (GO:MF), biological process (GO:BP), and cellular compartment (GO:CC) annotations, which are significantly enriched for the proteins in each cluster and computed using g:profiler. We also show the log(fold change) for genes in the cluster, which are more expressed in rumen tissue than on average across all tissues. For each module, nodes have been added in gray (proteins not in this cluster) if necessary to fully connect all nodes.

(A–C) We find 3 modules containing members of the PRD-SPRRII family, which are enriched for phosphate and mitochondrial metabolism (A and B) and regulation of cell growth mechanisms (C). We also find a module with TCHH-like two proteins enriched for immune response (D) and with S100-A2 and S100-A12 proteins enriched for transcriptional regulation and chromatin organization (E). The modules in (A), (B), and (E) are directly connected through *TARS* and *MRPL4*, which suggest a link between these functions in bovine rumen.

(F) We demonstrate that protein pairs with a predicted D-SCRIPT edge correspond to a significantly higher co-expression between their respective genes (one-sided Welch's t test). This co-expression signal gets even stronger when evaluated only on protein pairs in a functional module, suggesting that both the protein network and functional modules are biologically meaningful. See also Figures S4 and S5.

mouse S100-G, despite over an 81% sequence identity (Permyakov et al., 2020). Still, we hypothesize a connection to antimicrobial activity and innate immunity for proteins in cluster E.

We further identify two additional, smaller clusters. In one, a *PRD-SPRRII* homolog occurs in a module of 16 genes (cluster C, Figure 4C) that also contains *HOXC6* and *HOXA7* and is enriched for multicellular organism development, skeletal system development, and organ morphogenesis, suggesting a role for *PRD-SPRRII* in cell growth in the rumen. Other genes in this cluster include *SNIP*, with anti-inflammatory function (Shi et al., 2018), and several genes whose human homologs *FAP*, *PRRX1*, and *TP73* have been implicated in extracellular matrix remodeling and cancer metastasis (Guo et al., 2015; Yang et al., 2016; Rodríguez et al., 2018). Both of the *TCHHL2* homologs are part of the other small module (cluster D, Figure 4D), which has 10 genes and is enriched for B cell selection and proliferation of CD4-positive alpha-beta T cells, lymphocytes, and leukocytes, suggesting that *TCHHL2* plays a role in the immune response within the bovine rumen. The *TCHHL2* protein may be involved in cross-linking keratins at the ruminal surface (Garcia et al., 2017). The link between metabolism, cell growth, and the immune system is documented in Turner et al., and our analysis further suggests that these processes are involved in the modulation of immune response (Turner et al., 2016).

The earlier mentioned clusters contain many genes involved in rumen physiology and cow health, with some of their homologs implicated in human diseases, demonstrating the ability of D-SCRIPT to go directly from genome to functionally meaningful gene associations. However, similar to PPIs determined by *in vitro* assays, the predictions of D-SCRIPT should be cross-referenced with tissue-specific information when interpreting them in a particular tissue type. Consider cluster B, which consists predominantly of mitochondrial genes. Since D-SCRIPT has no knowledge of cellular compartments, this grouping has emerged naturally from the data, suggesting a biological signal. Indeed, we find many of these genes are highly expressed in the rumen (Figure 4B). On the other hand, 34 genes in cluster A are protein tyrosine phosphatases (PTPs), all with similar but not identical functions (Figure S4). It is possible that only some of these genes are involved in rumen biology, with the rest active in other tissues. RNA-seq data support this, identifying only a few PTPs as highly expressed in the rumen (Figure 4A). In general, the structure/function specificity of a protein family would help determine each member's tissue-specific selectivity. Interestingly, the large set of PTPs provides a natural setting to investigate D-SCRIPT's sensitivity to small sequence variations. PTP binding specificity is largely determined by the PTP catalytic signature motif (HCX_5R) (Kim and Ryu, 2012). We find that D-SCRIPT-predicted interaction probabilities (between SPRR-II and 29 PTP proteins) drop substantially when the entire eight residue motif is perturbed but remains high when only the less-conserved sites are randomized, indicating that D-SCRIPT is sensitive to residues that determine binding specificity (STAR Methods; Figure S5A). Further, we find that a systematic perturbation of each residue of the CDC14 subfamily of PTPs identifies the location of the catalytic signature motif completely *de novo*, which suggests that such an *in silico* mutagenesis experiment could be used to form hypotheses about binding mechanisms of uncharacterized proteins (STAR Methods; Figures S5B and S5C).

D-SCRIPT embeddings capture structure and interaction

One of our aims when designing D-SCRIPT was to capture the structural aspects of interaction—the per-protein embedding produced by the trained projection module should encode structural information. To examine this aspect, we randomly selected 300 proteins from the Protein Data Bank (PDB) and used ($n \times 100$)-dimensional D-SCRIPT embeddings of these proteins to predict protein structure. We randomly split the 300 PDB structures into a training set of 100 and a test set of 200, evaluating how well a logistic regression that uses the D-SCRIPT embeddings as the input predicts contacts between residues (STAR Methods). We show that a linear combination of features in the projection module output is able to recapitulate a significant subset of the ground-truth contacts, achieving a median per-structure AUPR of 0.19 over the test dataset (Figure 5). These results strongly suggest that the end-to-end training of D-SCRIPT—using only sequence data—results in an intermediate representation that captures structural information at the level of each protein.

We also sought to directly assess the utility of D-SCRIPT's embeddings for predicting PPIs by a nearest-neighbor search. We hypothesized that proteins that have similar embeddings are likely to interact in the same way; therefore, it is possible to find new interacting pairs by searching for proteins that are similar to known interacting pairs. We compared D-SCRIPT embeddings with several other protein sequence representations: a one-hot embedding categorizing each amino acid into one of seven classes based on biochemical properties ("AAClass," Shen et al., 2007), a 5-residue-context Skip-Gram embedding ("Vec5," Mikolov et al., 2013), a concatenation of Vec5 and AAClass (used in the input for PIPR), and a randomly generated fifty-dimensional embedding with values drawn uniformly from the range $[0, 1]$. Additionally, we used BLAST (Altschul et al., 1990) to search for neighbors of interacting proteins. We then evaluated the number of interacting pairs we found in the neighborhood of a small set of "seed pairs" and found that D-SCRIPT finds more interactions in the nearest neighbors of the seed pairs than all other embeddings in *H. sapiens* (Figure 5F), *S. cerevisiae* (Figure 5G), and *C. elegans* (Figure 5H).

D-SCRIPT contact maps recapitulate protein binding mechanisms

We investigated whether the interpretability of our model could aid in predicting inter-protein docking contacts. As an intermediate representation, the D-SCRIPT Contact Module (Figure 2) produces an inter-protein contact map, which predicts the probability of interaction between all pairs of residues in the candidate protein pair. We first sought to verify that the contact maps produced after training were consistent with our design goal: the maps corresponding to negative predictions should have uniformly near-zero contact probabilities, whereas those for positive predictions should be sparse but with isolated regions of high-probability contact prediction. We found that this was indeed the case generally and show some examples in Figure 6: the maximum predicted residue-contact probability is high for positive examples and low for negative examples.

Next, we assessed if the predicted contact map is physically representative of the actual docking mechanism of the interaction. We emphasize that this is a high bar given we do not provide any 3D information to the model nor any guidance on docking and, in principle, the model could perform well on the classification task without a physically accurate contact map. We performed this test using Hwang et al.'s benchmark dataset of docked protein structures (Hwang et al., 2010). For every pair of chains in each PDB complex in the benchmark set, we generated a candidate PPI. We applied our human-data-trained model on 295 candidate PPIs and evaluated the predicted contact maps against the ground-truth contacts (assessed at 8 Å).

In cases where our model predicted an interaction, we found the predicted contacts to indeed recapitulate the ground-truth contacts substantially (Figures 6A and 6B). Even in some of the cases where D-SCRIPT did not predict an interaction, the distribution of predicted contacts was nevertheless consistent with the ground truth (Figure 6C). To systematically evaluate the accuracy of the D-SCRIPT contact map, we evaluated the distance between the predicted and true contacts using an optimal transport metric and compared with a baseline established by randomly reshuffling the predicted matrix. We chose to measure similarity between regions of the two contact maps rather than measuring per-residue matches (with a metric such as binary

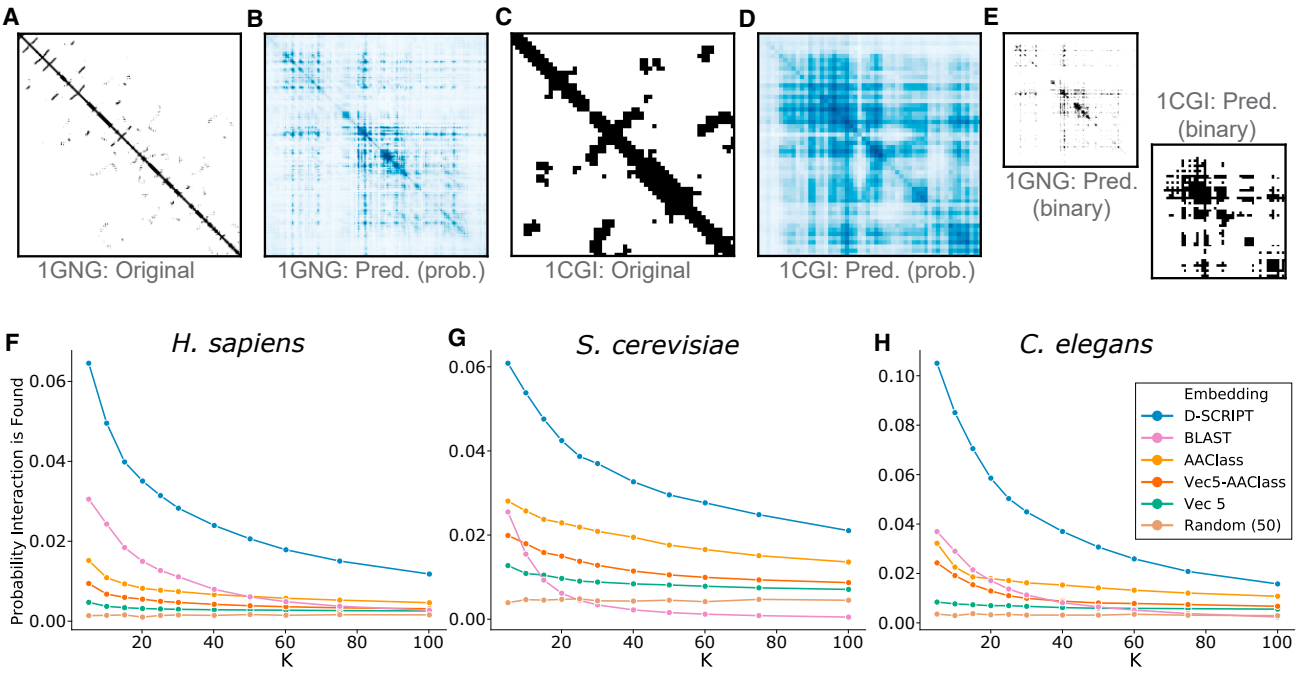


Figure 5. D-SCRIPT embeddings represent structure and interaction

After a full model has been trained to predict interaction, the low-dimensional embeddings learned by the projection module of D-SCRIPT can be used as meaningful representations of the protein in other applications.

(A–E) The PDB identifier 1GNG corresponds to a protein with 356 residues where the accuracy of using the D-SCRIPT embedding to predict self-contacts is near the median of cases we studied (AUPR = 0.19), whereas 1CGI corresponds to a short protein (54 residues) in which the embedding achieves a higher accuracy (AUPR = 0.38). On a set of 300 PDB structures, we assessed contacts at 8 Å (A and C) and, using a training set of 100 structures, trained a logistic regression to predict contacts (B and D) for the remaining structures. The binarization thresholds for (E) were chosen so as to result in the same number of contacts as in the original maps.

(F–H) D-SCRIPT embeddings also enable the accurate recovery of true interacting protein pairs in the neighborhood of known PPIs in human (F), yeast (G), and roundworm (H). D-SCRIPT embeddings recover more interacting proteins than any other embedding, regardless of species or number of neighbors checked. AAClass also performs well, likely because it characterizes biochemistry, which is preserved at longer evolutionary distances. BLAST performs well at low values of k but has difficulty recovering interactions for larger values—likely due to network rewiring over longer evolutionary distances.

cross-entropy) because the convolutional and max-pooling layers in our model aggregate over neighboring residues, thus diffusing the signal. We estimated the p value of the predicted contacts against 500 random trials, finding that in cases where D-SCRIPT predicted an interaction, the contact maps were substantially similar to the ground truth (median FDR-corrected q = 0.08, one-sided t test). Even in cases where D-SCRIPT did not predict an interaction, the similarity to the ground truth was higher than that of the random baselines (Figure 6E).

Performance

D-SCRIPT took approximately 3 days to train for 10 epochs on 843,602 training pairs and fits within a single 32GB GPU (i.e., graphics processing unit). Running time and GPU memory usage scales roughly quadratically, O(nm), with the protein lengths n,m, since D-SCRIPT models the full $n \times m$ contact map as an intermediate step. Prediction of new candidate pairs with a trained model is very fast, requiring on average 0.02 s/pair and less than 5GB of GPU memory. Since D-SCRIPT generalizes well across species, it needs to be trained only once on a large corpus of data and can be used to make predictions in a variety of settings.

DISCUSSION

We have introduced D-SCRIPT, an interpretable method for PPI prediction from sequence. D-SCRIPT pursues a structure-based approach, with the prediction score for a protein pair computed as the binding compatibility of their respective structures. Since structure is more conserved than sequence over evolutionary time (Ingles-Prieto et al., 2013), this physical model of interaction generalizes well across species. The intermediate contact map representation in the model is directly interpretable and can be used to validate the prediction or study the proteins' binding regions on a residue scale. D-SCRIPT thus joins the small but growing set of advances in interpretable deep-learning methods in computational biology (Hie et al., 2020; Luo et al., 2020a, 2020b). Our modular design additionally enables the investigation of model output at various stages, and we demonstrate that each layer captures incremental structural information.

The advantage of a sequence-based approach, such as D-SCRIPT, is that the input sequence data are almost always available, due to the enormous advances in low-cost genome sequencing. Compared with PIPR (Chen et al., 2019), the state-of-the-art deep-learning method that also takes sequences as inputs, D-SCRIPT generalizes better across species

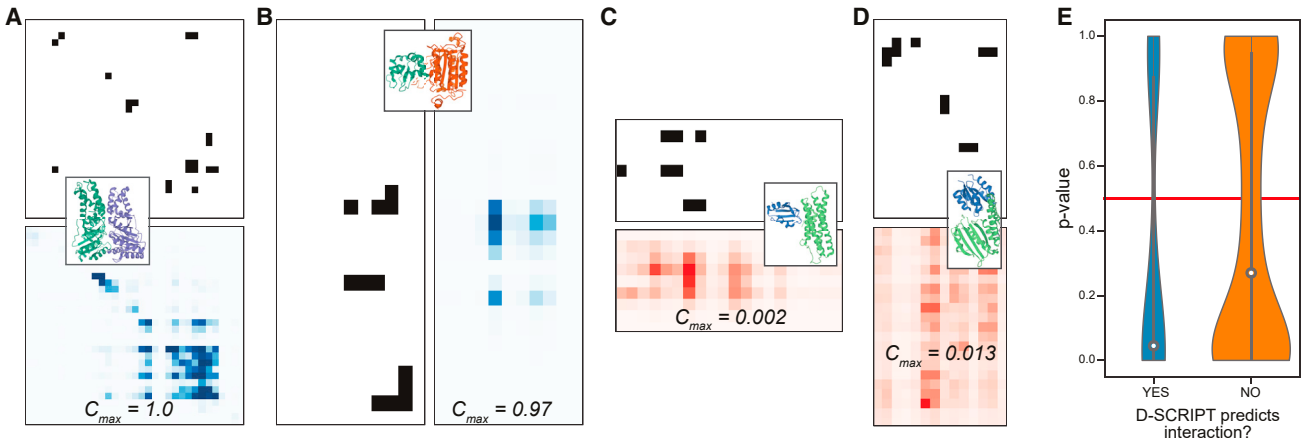


Figure 6. D-SCRIPT predicts biologically meaningful contact maps

We show inter-protein contact maps of protein structures known to dock together (Hwang et al., 2010).

(A–D) (A and B) correspond to pairs where D-SCRIPT correctly predicted an interaction, whereas (C and D) are cases where it incorrectly predicted no interaction. The black-and-white matrices correspond to the PDB ground truth, whereas the colored matrices correspond to D-SCRIPT's predicted contact map $\hat{C}$; for the latter, the color scales of (A and B) differ from (C and D). As designed, $\hat{C}$ is sparse but contains some large values in the case of positive predictions while its maximum C_{max} is very low for negative pairs.

(E) Shows a violin plot of a systematic evaluation (295 protein pairs, each with 500 bootstrap samples to generate the p value) of the 2D earth mover's distance-based similarity between $\hat{C}$ and the ground truth. Not only are the $\hat{C}$s of correctly predicted pairs substantially similar to the ground truth (median FDR-corrected q = 0.08, one-sided t test), even when D-SCRIPT incorrectly predicts that two proteins do not interact, its contact maps are still similar to ground truth. **PDB** identifiers: (A) 2J7P (A/D), (B) 1NW9 (B/A), (C) 3H2V (A/E), and (D) 1F51 (A/E).

and can, thus, be more effective for accurate *de novo* PPI predictions in non-model organisms or less-studied proteins in organisms such as the fly. We suspect that D-SCRIPT's relative success across species but under-performance on a within-species evaluation is due to the simplicity of the model and the extent to which it is regularized. These design choices enhance D-SCRIPT's generalizability, directing it to learn general structural aspects of the interaction rather than using network structure or the frequency of any individual protein as an interaction partner. However, for certain tasks, a balance between the cross-species generalizability of D-SCRIPT and the within-species specificity of other state-of-the-art methods may be desirable. A future research direction might be transfer learning to tune a pre-trained D-SCRIPT model to a target species, whereas another approach could be to integrate it with guilt-by-association graph-theoretic PPI predictions (Devkota et al., 2020).

Notably, D-SCRIPT does not require a multiple sequence alignment (MSA). However, the pre-trained language model used in D-SCRIPT was co-trained on MSAs over the entire protein corpus (Bepler and Berger, 2019, 2021), allowing its input featurization to implicitly capture some aspects of evolutionary conservation. Previously, co-evolution-based approaches that explicitly use MSAs have proven remarkably effective in reconstructing single-protein contact maps and 3D structures (Marks et al., 2011; Kamisetty et al., 2013; Liu et al., 2018). When extending them to PPI prediction, an additional challenge is to identify the correct correspondence order between the rows of the two MSAs. In prokaryotic genomes where synteny conservation can be very informative, methods such as ComplexContact (Zeng et al., 2018), EV Complex (Hopf et al., 2014; Green et al., 2021), and Gremlin (Cong et al., 2019) have been shown to perform well and provide residue-level interaction detail. However, there has been less success in extending these approaches to more complex, eukaryotic genomes. We found the need to compute MSAs to be a performance bottleneck (Dey et al., 2017), making it infeasible to perform eukaryotic genome-scale predictions with them and, therefore, limiting the applicability of an EV complex-like approach in our setting. Nonetheless, explicitly incorporating co-evolutionary insights could improve D-SCRIPT's accuracy, and future work might explore ways to do so without sacrificing speed. Insights from related advances in the prediction of contact maps and structures of individual proteins could also be incorporated into our model architecture (see STAR Methods for additional discussion).

D-SCRIPT illustrates that learning the language of individual proteins, a remarkably successful deep-learning effort, also helps decode the language of protein interactions. We leverage Bepler and Berger's pre-trained language model, allowing us to indirectly benefit from the rich data on 3D structures of individual proteins (Bepler and Berger, 2019). In contrast, a PPI prediction method that was directly supervised with 3D structures of protein complexes, in order to learn the physical mechanism of interaction, would need to contend with the relatively small size of that corpus (Singh et al., 2006; Singh et al., 2010; Hosur et al., 2011).

There is a pressing need for scalable computational methods to infer a gene's function from its sequence in non-model organisms. Although the sequencing revolution has helped make genomes more widely available, there remains a dearth of functional data. PPI prediction with D-SCRIPT is fast, making genome-scale screening feasible. For instance, we were able to evaluate 50 million candidate PPIs in *B. taurus* in 8 days on a single GPU. With D-SCRIPT, a workflow consisting of genome-scale PPI prediction, followed by graph-theoretic analysis of the PPI network to identify functional modules, can

generate high-confidence predictions of gene function at scale; we demonstrated this in our cow rumen case study. Such *de novo* PPI prediction can be useful even in model organisms, such as *C. elegans*, for which the known portion of the PPI network is still quite sparse. In other organisms (e.g., *D. melanogaster*) where some PPI data do exist, future work could productively combine those data with D-SCRIPT predictions. We hope that its combination of broad applicability, cross-species accuracy, and speed will make D-SCRIPT a useful community resource for addressing the "genome-to-phenome" challenge.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- METHOD DETAILS
  - Data sources
  - Model architecture
  - Training
  - Analysis
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cels.2021.08.010.

ACKNOWLEDGMENTS

We thank Tristan Bepler for helpful discussions and technical assistance. S.S., R.S., and B.B. were partially supported by the NIH grants R01-GM081871 and R35-GM141861. L.C. was supported by NSF grant OAC-1939263.

AUTHOR CONTRIBUTIONS

All authors conceived of the project, developed the methods, interpreted the results, and wrote the manuscript. S.S wrote the software, with inputs from R.S.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: April 2, 2021
Revised: August 1, 2021
Accepted: August 19, 2021
Published: September 17, 2021

REFERENCES

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., et al. (2000). The genome sequence of Drosophila melanogaster. Science *287*, 2185–2195.

Alborzi, S.Z., Ritchie, D.W., and Devignes, M.D. (2018). Computational discovery of direct associations between GO terms and protein domains. BMC Bioinformatics *19*, 413. https://doi.org/10.1186/s12859-018-2380-2.

Alonso, A., and Pulido, R. (2016). The extended human PTP ome: a growing tyrosine phosphatase family. FEBS Journal *283*, 1404–1429.

Alonso, A., Sasin, J., Bottini, N., Friedberg, I., Friedberg, I., Osterman, A., Godzik, A., Hunter, T., Dixon, J., and Mustelin, T. (2004). Protein tyrosine phosphatases in the human genome. Cell *117*, 699–711.

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

Aschenbach, J.R., Kristensen, N.B., Donkin, S.S., Hammon, H.M., and Penner, G.B. (2010). Gluconeogenesis in dairy cows: the secret of making sweet milk from sour dough. IUBMB Life *62*, 869–877.

Bepler, T., and Berger, B. (2019). Learning protein sequence embeddings using information from structure. arXiv https://arxiv.org/abs/1902.08661.

Bepler, T., and Berger, B. (2021). Learning the protein language: evolution, structure, and function. Cell Syst *12*, 654–669.e3. https://doi.org/10.1016/j.cels.2021.05.017.

Besse, A., Wu, P., Bruni, F., Donti, T., Graham, B.H., Craigen, W.J., McFarland, R., Moretti, P., Lalani, S., Scott, K.L., et al. (2015). The GABA transaminase, ABAT, is essential for mitochondrial nucleoside metabolism. Cell Metab *21*, 417–427.

Brauksiepe, B., Mujica, A.O., Herrmann, H., and Schmidt, E.R. (2008). The serine/threonine kinase Stk33 exhibits autophosphorylation and phosphorylates the intermediate filament protein vimentin. BMC Biochem *9*, 25.

Cao, M., Pietras, C.M., Feng, X., Doroschak, K.J., Schaffner, T., Park, J., Zhang, H., Cowen, L.J., and Hescott, B.J. (2014). New directions for diffusion-based network prediction of protein function: incorporating pathways with confidence. Bioinformatics *30*, i219–i227.

Cao, M., Zhang, H., Park, J., Daniels, N.M., Crovella, M.E., Cowen, L.J., and Hescott, B. (2013). Going the distance for protein function prediction: a new distance metric for protein interaction networks. PLoS One *8*, e76339.

Castro, L.F.C., Lopes-Marques, M., Wilson, J.M., Rocha, E., Reis-Henriques, M.A., Santos, M.M., and Cunha, I. (2012). A novel acetyl-CoA synthetase short-chain subfamily member 1 (Acss1) gene indicates a dynamic history of paralogue retention and loss in vertebrates. Gene *497*, 249–255.

Chen, M., Ju, C.J., Zhou, G., Chen, X., Zhang, T., Chang, K.W., Zaniolo, C., and Wang, W. (2019). Multifaceted protein-protein interaction prediction based on Siamese residual RCNN. Bioinformatics *35*, i305–i314. https://doi.org/10.1093/bioinformatics/btz328.

Chen, Y., Ruan, Z.R., Wang, Y., Huang, Q., Xue, M.Q., Zhou, X.L., and Wang, E.D. (2018). A threonyl-tRNA synthetase-like protein has tRNA aminoacylation and editing activities. Nucleic Acids Res *46*, 3643–3656.

Cho, H., Berger, B., and Peng, J. (2016). Compact integration of multi-network topology for functional analysis of genes. Cell Syst *3*, 540–548.e5. https://doi.org/10.1016/j.cels.2016.10.017.

Choobdar, S., Ahsen, M.E., Crawford, J., Tomasoni, M., Fang, T., Lamparter, D., Lin, J., Hescott, B., Hu, X., Mercer, J., et al. (2019). Assessment of network module identification across complex diseases. Nat. Methods *16*, 843–852.

Cong, Q., Anishchenko, I., Ovchinnikov, S., and Baker, D. (2019). Protein interaction networks revealed by proteome coevolution. Science *365*, 185–189.

Cowen, L., Ideker, T., Raphael, B.J., and Sharan, R. (2017). Network propagation: a universal amplifier of genetic associations. Nat. Rev. Genet. *18*, 551–562.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res *14*, 1188–1190. https://doi.org/10.1101/gr.849004.

Devkota, K., Murphy, J.M., and Cowen, L.J. (2020). GLIDE: combining local methods and diffusion state embeddings to predict missing interactions in biological networks. Bioinformatics *36*, i464–i473.

Dey, A., Saha, I., and Maulik, U. (2017). A survey on multiple sequence alignment using metaheuristics. In 7th International Conference on Communication Systems and Network Technologies (CSNT)), pp. 279–284. https://doi.org/10.1109/CSNT.2017.8418552.

Dodson, M.V., Hausman, G.J., Guan, L., Du, M., Rasmussen, T.P., Poulos, S.P., Mir, P., Bergen, W.G., Fernyhough, M.E., McFarland, D.C., et al. (2010). Lipid metabolism, adipocyte depot physiology and utilization of meat

animals as experimental models for metabolic research. Int. J. Biol. Sci. *6*, 691–699.

Eddy, S.R. (2009). A new generation of homology search tools based on probabilistic inference. In Genome Informatics 2009 (World Scientific), pp. 205–211. https://doi.org/10.1142/9781848165632_0019.

Edgar, R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics *5*, 113.

Fassone, E., Duncan, A.J., Taanman, J., Pagnamenta, A.T., Sadowski, M.I., Holand, T., Qasim, W., Rutland, P., Calvo, S.E., Mootha, V.K., et al. (2010). FOXRED1, encoding an FAD-dependent oxidoreductase complex-I-specific molecular chaperone, is mutated in infantile-onset mitochondrial encephalopathy. Hum. Mol. Genet. *19*, 4837–4847.

Fields, S., and Song, O. (1989). A novel genetic system to detect protein–protein interactions. Nature *340*, 245–246.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., et al. (2014). Pfam: the protein families database. Nucleic Acids Res *42*, D222–D230. https://doi.org/10.1093/nar/gkt1223.

Flamary, R., and Courty, N. (2017). POT python optimal transport library. https://pythonot.github.io/.

Franz, M., Rodriguez, H., Lopes, C., Zuberi, K., Montojo, J., Bader, G.D., and Morris, Q. (2018). GeneMANIA update 2018. Nucleic Acids Res *46*, W60–W64.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152.

Gao, X., and Oba, M. (2016). Characteristics of dairy cows with a greater or lower risk of subacute ruminal acidosis: volatile fatty acid absorption, rumen digestion, and expression of genes in rumen epithelial cells. J. Dairy Sci. *99*, 8733–8745.

Garcia, M., Bradford, B.J., and Nagaraja, T.G. (2017). Invited review: ruminal microbes, microbial products, and systemic inflammation. Prof. Anim. Sci. *33*, 635–650.

Ghezzi, D., Baruffini, E., Haack, T.B., Invernizzi, F., Melchionda, L., Dallabona, C., Strom, T.M., Parini, R., Burlina, A.B., Meitinger, T., et al. (2012). Mutations of the mitochondrial-tRNA modifier MTO1 cause hypertrophic cardiomyopathy and lactic acidosis. Am. J. Hum. Genet. *90*, 1079–1087.

Ginguay, A., Cynober, L., Curis, E., and Nicolis, I. (2017). Ornithine aminotransferase, an important glutamate-metabolizing enzyme at the crossroads of multiple metabolic pathways. Biology *6*, 18.

Gohla, A. (2019). Do metabolic HAD phosphatases moonlight as protein phosphatases? Biochim. Biophys. Acta Mol. Cell Res. *1866*, 153–166.

Gordon, D.E., Jang, G.M., Bouhaddou, M., Xu, J., Obernier, K., White, K.M., O'Meara, M.J., Rezelj, V.V., Guo, J.Z., Swaney, D.L., et al. (2020). A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature *583*, 459–468.

Green, A.G., Elhabashy, H., Brock, K.P., Maddamsetti, R., Kohlbacher, O., and Marks, D.S. (2021). Large-scale discovery of protein interactions at residue resolution using co-evolution calculated from genomic sequences. Nat. Commun. *12*, 1396. https://doi.org/10.1038/s41467-021-21636-z.

Guo, J., Fu, Z., Wei, J., Lu, W., Feng, J., and Zhang, S. (2015). PRRX1 promotes epithelial–mesenchymal transition through the Wnt/β-catenin pathway in gastric cancer. Med. Oncol. *32*, 393.

Halstead, M.M., Islas-Trejo, A., Goszczynski, D.E., Medrano, J.F., Zhou, H., and Ross, P.J. (2021). Large-scale multiplexing permits full-length transcriptome annotation of 32 bovine tissues from a single nanopore flow cell. Front. Genet. *12*, 664260. https://doi.org/10.3389/fgene.2021.664260.

Hashemifar, S., Neyshabur, B., Khan, A.A., and Xu, J. (2018). Predicting protein-protein interactions through sequence-based deep learning. Bioinformatics *34*, i802–i810. https://doi.org/10.1093/bioinformatics/bty573.

Heizmann, C.W. (2019). S100 proteins: diagnostic and prognostic biomarkers in laboratory medicine. Biochim. Biophys. Acta Mol. Cell Res. *1866*, 1197–1206.

Hie, B., Bryson, B.D., and Berger, B.A. (2020). Leveraging uncertainty in machine learning accelerates biological discovery and design. Cell Syst *11*, 461–477.e9.

Hopf, T.A., Schärfe, C.P., Rodrigues, J.P., Green, A.G., Kohlbacher, O., Sander, C., Bonvin, A.M., and Marks, D.S. (2014). Sequence co-evolution gives 3D contacts and structures of protein complexes. eLife *3*, e03430. https://doi.org/10.7554/eLife.03430.

Hosur, R., Peng, J., Vinayagam, A., Stelzl, U., Xu, J., Perrimon, N., Bienkowska, J., and Berger, B. (2012). A computational framework for boosting confidence in high-throughput protein-protein interaction datasets. Genome Biol *13*, R76.

Hosur, R., Xu, J., Bienkowska, J., and Berger, B. (2011). iWRAP: an interface threading approach with application to prediction of cancer-related protein–protein interactions. J. Mol. Biol. *405*, 1295–1310.

Hsu, K., Champaiboon, C., Guenther, B.D., Sorenson, B.S., Khammanivong, A., Ross, K.F., Geczy, C.L., and Herzberg, M.C. (2009). Anti-infective protective properties of S100 calgranulins. Antiinflamm. Antiallergy. Agents Med. Chem. *8*, 290–305.

Huang, H., Cao, J., Guo, G., Li, X., Wang, Y., Yu, Y., Zhang, S., Zhang, Q., and Zhang, Y. (2019). Genome-wide association study identifies QTLs for displacement of abomasum in Chinese Holstein cattle1. J. Anim. Sci. *97*, 1133–1142.

Hulovatyy, Y., Solava, R.W., and Milenković, T. (2014). Revealing missing parts of the interactome via link prediction. PLoS One *9*, e90073.

Hwang, H., Vreven, T., Janin, J., and Weng, Z. (2010). Protein–protein docking benchmark version 4.0. Proteins *78*, 3111–3114.

Ingles-Prieto, A., Ibarra-Molero, B., Delgado-Delgado, A., Perez-Jimenez, R., Fernandez, J.M., Gaucher, E.A., Sanchez-Ruiz, J.M., and Gavira, J.A. (2013). Conservation of protein structure over four billion years. Structure *21*, 1690–1697.

Jiang, Y., Xie, M., Chen, W., Talbot, R., Maddox, J.F., Faraut, T., Wu, C., Muzny, D.M., Li, Y., Zhang, W., et al. (2014). The sheep genome illuminates biology of the rumen and lipid metabolism. Science *344*, 1168–1173. https://doi.org/10.1126/science.1252806.

Jing, L., Yang, C., and Guiying, L. (2009). Relationship between the polymorphisms of DECR1 gene and meat quality traits in Yanbian yellow cattle. J. Anhui Agric. Sci. *34*, 20.

Kamiński, S., Brym, P., and Wójcik, E. (2009). A note on associations between polymorphism within the 2,4-dienoyl-CoA reductase gene (DECR1) and growth rate of Polish Landrace boars. J. Anim. Feed Sci. *18*, 71–77.

Kamisetty, H., Ovchinnikov, S., and Baker, D. (2013). Assessing the utility of coevolution-based residue–residue contact predictions in a sequence- and structure-rich era. Proc. Natl. Acad. Sci. USA *110*, 15674–15679.

Kim, S.J., and Ryu, S.E. (2012). Structure and catalytic mechanism of human protein tyrosine phosphatome. BMB Rep *45*, 693–699.

Kovács, I.A., Luck, K., Spirohn, K., Wang, Y., Pollis, C., Schlabach, S., Bian, W., Kim, D.K., Kishore, N., Hao, T., et al. (2019). Network-based prediction of protein interactions. Nat. Commun. *10*, 1240. https://doi.org/10.1038/s41467-019-09177-y.

Kozlyuk, N., Monteith, A.J., Garcia, V., Damo, S.M., Skaar, E.P., and Chazin, W.J. (2019). S100 proteins in the innate immune response to pathogens. Methods Mol. Biol. *1929*, 275–290.

Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P., et al. (2006). Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature *440*, 637–643.

Kumar, A., and Snyder, M. (2002). Protein complexes take the bait. Nature *415*, 123–124.

Lee, N., Kim, D.K., Han, S.H., Ryu, H.G., Park, S.J., Kim, K.T., and Choi, K.Y. (2017). Comparative interactomes of VRK1 and VRK3 with their distinct roles in the cell cycle of liver cancer. Mol. Cells *40*, 621–631.

Lei, C., and Ruan, J. (2013). A novel link prediction algorithm for reconstructing protein–protein interaction networks by topological similarity. Bioinformatics *29*, 355–364.

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics *22*, 1658–1659.

Liu, Y., Palmedo, P., Ye, Q., Berger, B., and Peng, J. (2018). Enhancing evolutionary couplings with deep convolutional neural networks. Cell Syst *6*, 65–74.e3.

Luo, Y., Peng, J., and Ma, J. (2020a). When causal inference meets deep learning. Nat. Mach. Intell. *2*, 426–427.

Luo, Y., Vo, L., Ding, H., Su, Y., Liu, Y., Qian, W.W., Zhao, H., and Peng, J. (2020b). Evolutionary context-integrated deep sequence modeling for protein engineering. Lecture Notes in Computer Science, 261–263.

Marks, D.S., Colwell, L.J., Sheridan, R., Hopf, T.A., Pagnani, A., Zecchina, R., and Sander, C. (2011). Protein 3D structure computed from evolutionary sequence variation. PLoS One *6*, e28766. https://doi.org/10.1371/journal.pone.0028766.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Adv. Neural Inf. Process. Syst. 3111–3119.

Mutowo, P., Bento, A.P., Dedman, N., Gaulton, A., Hersey, A., Lomax, J., and Overington, J.P. (2016). A drug target slim: using gene ontology and gene ontology annotations to navigate protein-ligand target space in ChEMBL. J. Biomed. Semantics *7*, 59.

Navlakha, S., and Kingsford, C. (2010). The power of protein interaction networks for associating genes with diseases. Bioinformatics *26*, 1057–1063.

Permyakov, S.E., Yundina, E.N., Kazakov, A.S., Permyakova, M.E., Uversky, V.N., and Permyakov, E.A. (2020). Mouse S100G protein exhibits properties characteristic of a calcium sensor. Cell Calcium *87*, 102185.

Pohjoismäki, J.L., Wanrooij, S., Hyvärinen, A.K., Goffart, S., Holt, I.J., Spelbrink, J.N., and Jacobs, H.T. (2006). Alterations to the expression level of mitochondrial transcription factor A, TFAM, modify the mode of mitochondrial DNA replication in cultured human cells. Nucleic Acids Res *34*, 5815–5828.

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., and Vilo, J. (2019). g:profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res *47*, W191–W198. https://doi.org/10.1093/nar/gkz369.

Reiter, V., Matschkal, D.M., Wagner, M., Globisch, D., Kneuttinger, A.C., Müller, M., and Carell, T. (2012). The CDK5 repressor CDK5RAP1 is a methylthiotransferase acting on nuclear and mitochondrial RNA. Nucleic Acids Res *40*, 6235–6240.

Richoux, F., Servantie, C., Borès, C., and Téletchéa, S. (2019). Comparing two deep learning sequence-based models for protein-protein interaction prediction. http://arxiv.org/abs/1901.06268.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C.L., and Ma, J. (2019). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. bioRxiv. https://doi.org/10.1101/622803.

Rodionov, R.N., Jarzebska, N., Weiss, N., and Lentz, S.R. (2014). AGXT2: a promiscuous aminotransferase. Trends Pharmacol. Sci. *35*, 575–582.

Rodríguez, N., Peláez, A., Barderas, R., and Domínguez, G. (2018). Clinical implications of the deregulated TP73 isoforms expression in cancer. Clin. Transl. Oncol. *20*, 827–836.

Sahni, N., Yi, S., Taipale, M., Fuxman Bass, J.I., Coulombe-Huntington, J., Yang, F., Peng, J., Weile, J., Karras, G.I., Wang, Y., et al. (2015). Widespread macromolecular interaction perturbations in human genetic disorders. Cell *161*, 647–660.

Sasaki, K., Yamagishi, N., Kizaki, K., Sasaki, K., Devkota, B., and Hashizume, K. (2014). Microarray-based gene expression profiling of peripheral blood mononuclear cells in dairy cows with experimental hypocalcemia and milk fever. J. Dairy Sci. *97*, 247–258.

Schneider, T.D., and Stephens, R.M. (1990). Sequence logos: a new way to display consensus sequences. Nucleic Acids Res *18*, 6097–6100. https://doi.org/10.1093/nar/18.20.6097.

Schwarzenbacher, H., Burgstaller, J., Seefried, F.R., Wurmser, C., Hilbe, M., Jung, S., Fuerst, C., Dinhopl, N., Weissenböck, H., Fuerst-Waltl, B., et al. (2016). A missense mutation in TUBD1 is associated with high juvenile mortality in Braunvieh and Fleckvieh cattle. BMC Genomics *17*, 400.

Sharan, R., Ulitsky, I., and Shamir, R. (2007). Network-based prediction of protein function. Mol. Syst. Biol. *3*, 88.

Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., Li, Y., and Jiang, H. (2007). Predicting protein–protein interactions based only on sequences information. Proc. Natl. Acad. Sci. USA *104*, 4337–4341.

Shi, Y., He, C., Ma, C., Yu, T., Cong, Y., Cai, W., and Liu, Z. (2018). Smad nuclear interacting protein 1 (SNIP1) inhibits intestinal inflammation through regulation of epithelial barrier function. Mucosal Immunol *11*, 835–845.

Singh, R., Xu, J., and Berger, B. (2006). Struct2Net: integrating structure into protein-protein interaction prediction. In Proceedings of the Pacific Symposium on bBocomputing 2006 (World Scientific)), pp. 403–414. https://doi.org/10.1142/9789812701626_0037.

Singh, R., Park, D., Xu, J., Hosur, R., and Berger, B. (2010). Struct2Net: a web service to predict protein–protein interactions using a structure-based approach. Nucleic Acids Res *38*, W508–W515.

Soma, S., Morgada, M.N., Naik, M.T., Boulet, A., Roesler, A.A., Dziuba, N., Ghosh, A., Yu, Q., Lindahl, P.A., Ames, J.B., et al. (2019). COA6 is structurally tuned to function as a thiol-disulfide oxidoreductase in copper delivery to mitochondrial cytochrome c oxidase. Cell Rep *29*, 4114–4126.e5.

Sonnhammer, E.L.L., Eddy, S.R., and Durbin, R. (1997). Pfam: A comprehensive database of protein domain families based on seed alignments. Proteins *28*, 405–420.

Srivastava, A., Srivastava, K.R., Hebbar, M., Galada, C., Kadavigrere, R., Su, F., Cao, X., Chinnaiyan, A.M., Girisha, K.M., Shukla, A., and Bielas, S.L. (2018). Genetic diversity of NDUFV1-dependent mitochondrial complex I deficiency. Eur. J. Hum. Genet. *26*, 1582–1587.

Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., et al. (2019). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res *47*, D607–D613.

Tai, K.S., Socher, R., and Manning, C.D. (2015). Improved semantic representations from tree-structured long short-term memory networks. arXiv, arXiv:1503.00075.

Taipale, M., Tucker, G., Peng, J., Krykbaeva, I., Lin, Z.Y., Larsen, B., Choi, H., Berger, B., Gingras, A.C., and Lindquist, S. (2014). A quantitative chaperone interaction network reveals the architecture of cellular protein homeostasis pathways. Cell *158*, 434–448.

Thurmond, J., Goodman, J.L., Strelets, V.B., Attrill, H., Gramates, L.S., Marygold, S.J., Matthews, B.B., Millburn, G., Antonazzo, G., Trovisco, V., et al. (2019). FlyBase 2.0: the next generation. Nucleic Acids Res *47*, D759–D765.

Turner, M.L., Cronin, J.G., Noleto, P.G., and Sheldon, I.M. (2016). Glucose availability and AMP-activated protein kinase link energy metabolism and innate immunity in the bovine endometrium. PLoS One *11*, e0151416.

Wang, B., Pourshafeie, A., Zitnik, M., Zhu, J., Bustamante, C.D., Batzoglou, S., and Leskovec, J. (2018). Network enhancement as a general method to denoise weighted biological networks. Nat. Commun. *9*, 3108. https://doi.org/10.1038/s41467-018-05469-x.

Wei, X., Zhang, Y., Fu, Z., and Zhang, L. (2013). The association between polymorphisms in the MRPL4 and TNF-α genes and susceptibility to allergic rhinitis. PLOS One *8*, e57981.

Yadav, P., Kumar, P., Mukesh, M., Kataria, R.S., Yadav, A., Mohanty, A.K., and Mishra, B.P. (2015). Kinetics of lipogenic genes expression in milk purified mammary epithelial cells (MEC) across lactation and their correlation with milk and fat yield in buffalo. Res. Vet. Sci. *99*, 129–136.

Yang, X., Lin, Y., Shi, Y., Li, B., Liu, W., Yin, W., Dang, Y., Chu, Y., Fan, J., and He, R. (2016). FAP promotes immunosuppression by cancer-associated fibroblasts in the tumor microenvironment via STAT3–CCL2 signaling. Cancer Res *76*, 4124–4135.

Yu, G., Wang, L.G., Han, Y., and He, Q.Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS A J. Integr. Biol. *16*, 284–287.

Yu, J.T., Liu, Y., Dong, P., Cheng, R.E., Ke, S.X., Chen, K.Q., Wang, J.J., Shen, Z.S., Tang, Q.Y., and Zhang, Z. (2019). Up-regulation of antioxidative proteins Trx1, TXNL1 and TXNRD1 in the cortex of PTZ kindling seizure model mice. PLoS One *14*, e0210670.

Zarrin, M., Wellnitz, O., van Dorland, H.A., Gross, J.J., and Bruckmaier, R.M. (2014). Hyperketonemia during lipopolysaccharide-induced mastitis affects systemic and local intramammary metabolism in dairy cows. J. Dairy Sci. *97*, 3531–3541.

Zeng, H., Wang, S., Zhou, T., Zhao, F., Li, X., Wu, Q., and Xu, J. (2018). ComplexContact: a web server for inter-protein contact prediction using deep learning. Nucleic Acids Res *46*, W432–W437. https://doi.org/10.1093/nar/gky420.

Zhao, C., and Wang, Z. (2018). GOGO: an improved algorithm to measure the semantic similarity between gene ontology terms. Sci. Rep. *8*, 15107.

Zhao, J.M., and Qi, T.G. (2021). The role of TXNL1 in disease: treatment strategies for cancer and diseases with oxidative stress. Mol. Biol. Rep. *48*, 2929–2934.

Zhu, Z., and Leung, G.K.K. (2020). More than a metabolic enzyme: MTHFD2 as a novel target for anticancer therapy? Front. Oncol. *10*, 658.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
Protein sequence data	Szklarczyk et al., 2019	https://string-db.org/; RRID:SCR_005223
Protein interaction data	Szklarczyk et al., 2019	https://string-db.org; RRID:SCR_005223
Docking benchmark dataset	Hwang et al., 2010	https://zlab.umassmed.edu/benchmark/
SARS-CoV-2 sequences and interactions	Gordon et al., 2020	https://www.ebi.ac.uk/pride/archive/projects/PXD018117
ARS-UCD1.2 *Bos taurus* genome assembly	NCBI Assembly	https://www.ncbi.nlm.nih.gov/assembly/GCF_002263795.1/
Bos taurus RNA-seq	Halstead et al. 2021	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE160028; GEO: GSE160028
Processed data, trained models, and new predictions	This study	https://doi.org/10.5281/zenodo.5140612
Software and Algorithms		
D-SCRIPT	This study	http://dscript.csail.mit.edu/ (https://doi.org/10.5281/zenodo.5140508)
Protein sequence embeddings	Bepler and Berger, 2019	https://github.com/tbepler/protein-sequence-embedding-iclr2019
PIPR	Chen et al., 2019	https://github.com/muhaochen/seq_ppi
DeepPPI	Richoux et al., 2019	https://gitlab.univ-nantes.fr/richoux-f/DeepPPI
DSD	Cao et al. 2014	https://dsd.cs.tufts.edu/capdsd
CD-HIT	Fu et al., 2012	http://weizhongli-lab.org/cd-hit/
GOGO	Zhao and Wang, 2018	http://dna.cs.miami.edu/GOGO/
POT Python Optimal Transport	Flamary and Courty, 2017	https://pythonot.github.io/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for materials and code should be directed to and will be fulfilled by the Lead Contact, Bonnie Berger (bab@mit.edu).

Materials availability

This study did not generate new materials.

Data and code availability

- Sequence data, PPI training data, predicted edges, and trained models have been deposited at Zenodo at https://doi.org/10.5281/zenodo.5140612 and are publicly available as of the date of publication. This paper analyzes existing, publicly available data from Szklarczyk et al., 2019, Hwang et al., 2010, Gordon et al., 2020, and the NCBI. The accession numbers for all datasets are listed in the key resources table.
- All original code is publicly available at http://dscript.csail.mit.edu, has been deposited at Zenodo at https://doi.org/10.5281/zenodo.5140508, and is publicly available as of the date of publication.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Data sources

PPI data set

To evaluate the performance of D-SCRIPT in predicting protein-protein interactions, we use data from the STRING database (version 11) (Szklarczyk et al., 2019). STRING contains protein pairs corresponding to a variety of primary sources and interaction modalities

(e.g., binding vs co-expression). In order to select only high-confidence physical protein interactions, we limited our positive examples to binding interactions associated with a positive experimental-evidence score. From this set, we removed PPIs involving very short proteins (shorter than 50 amino acids) and, due to GPU memory constraints, also excluded proteins longer than 800 amino acids. Next, we removed PPIs with high sequence redundancy to other PPIs, following the precedent of previous approaches (Hashemifar et al., 2018; Chen et al., 2019). Specifically, we clustered proteins at the 40% similarity threshold using CD-HIT (Li and Godzik, 2006; Fu et al., 2012), and a PPI (A-B) was considered sequence redundant (and excluded) if we had already selected another PPI (C-D) such that the protein pairs (A, C) and (B, D) each shared a CD-HIT cluster. Removing sequence redundant PPIs from the data set prevents the model from memorizing interactions based on sequence similarity alone. To generate negative examples of PPI, we followed (Hashemifar et al., 2018) and randomly paired proteins from the non-redundant set, choosing a 10:1 negative-to-positive ratio to reflect the intuition that true positive PPIs are likely rare. Our human PPI data set contained 47,932 positive and 479,320 negative protein interactions, of which we set apart 80% (i.e. 38,345 positive examples) for training and 20% (i.e. 9,587 positive examples) for validation. For each of 5 model organisms (Table 1) we selected 5,000 positive interactions and 50,000 negative interactions using this procedure, with the exception of *E.coli* (2,000/20,000) where the available set of positive examples in STRING was limited.

BioGRID interaction data

We sourced the PPI data in Figure 1 from BioGRID. While we have sourced PPI data from the STRING database everywhere else in this paper, here we chose to use BioGRID because the publication date of a PPI is easily accessible in BioGRID, allowing us to estimate the number of PPIs assayed in the last five years. While the BioGRID selection may not precisely match the STRING selection due to curation differences between the two databases, our primary aim here is conveying the relative data availability across species; this estimate should not be substantially impacted by differences in curation.

Model architecture

Input & structure-aware embedding

The input to D-SCRIPT is a pair of protein sequences S_1, S_2 with lengths (n, m) and it outputs an interaction probability $\hat{p} \in [0,1]$ and a predicted-contact matrix $\hat{C} \in [0,1]^{n \times m}$. We first generate embeddings $E_1 \in \mathcal{R}^{n \times d_0}, E_2 \in \mathcal{R}^{m \times d_0}$ by embedding S_1 and S_2 with a pre-trained model from Bepler and Berger. Their model is a Bi-LSTM (bidirectional long short-term memory) neural network trained on three independent pieces of information: 1) the protein's SCOP classification, indicating its general structure, 2) self-contact map of a protein's 3-D structure, and 3) sequence alignment of similar proteins. These embeddings capture both local and global structural features of the protein sequences: the d_0-dimensional encoding of each amino acid contains information not just about the amino acid and its immediate context, but also the global structure of the protein. This is a key distinction from other approaches (e.g. Chen et al.'s in PIPR), where each amino acid's embedding represents just its biochemical properties or a short-range context (e.g., 5-7 residues) around it. We note that alternative embeddings (Rives et al., 2019; Luo et al., 2020a, 2020b) can potentially be substituted here.

Projection module

In the projection module, we reduce E_1 and E_2 to d-dimensional representations using a fully-connected linear layer (multi-layer perceptron) with d_0 input and d output nodes. Specifically, given an input embedding $E \in \mathcal{R}^{n \times d_0}$, we compute the embedding projection $Z \in \mathcal{R}^{n \times d}_{\geq 0}$ as

$$Z_i = \text{Drop}(\text{ReLU}(E_i W_1 + b_1)) \ \forall i \in 1 \ldots n \quad \text{(Equation 1)}$$

with $W_1 \in \mathcal{R}^{d_0 \times d}, b_1 \in \mathcal{R}^d$ as learned weights and biases. The rectified linear unit (ReLU) is a non-linear operation which applies the transformation $\text{ReLU}(x) = \max(0, x)$. The dropout layer (Drop) randomly sets 50% of the weights to zero, helping prevent over-fitting in W_1.

Residue contact module

The residue contact model takes the d-dimensional embeddings Z_1, Z_2 and models the interaction between the residues of each protein. First, for each pair of residue embeddings $Z_{1_i}, Z_{2_j} \in \mathcal{R}^d, i \in 1, \ldots, n, j \in 1, \ldots, m$, we compute a broadcast matrix with hidden dimension h, $B_{z_0, z_1} \in \mathcal{R}^{n \times m \times h}_{\geq 0}$

$$\text{diff}_{i,j} = \left| Z_{0_i} \ominus Z_{1_j} \right| \quad \text{(Equation 2)}$$

$$\text{mul}_{i,j} = Z_{0_i} \odot Z_{1_j} \quad \text{(Equation 3)}$$

$$B_{i,j} = \text{ReLU}(\text{Batch}([\text{diff}_{i,j}, \text{mul}_{i,j}] W_2 + b_2)) \quad \text{(Equation 4)}$$

where $\ominus$ indicates the element-wise difference and $\odot$ indicates the Hadamard product. This featurization is symmetric and has been previously used in natural language processing (NLP) and protein sequence modeling tasks (Tai et al., 2015; Bepler and Berger, 2019).

$W_2 \in \mathcal{R}^{2d \times h}, b_2 \in \mathcal{R}^h$ are the learned weights and biases. The batch normalization operation normalizes the mean and variance of the input features, thus stabilizing the learning process. Each element $B_{i,j}$ captures the direct interaction between residues S_{1_i} and S_{2_j}. The broadcast matrix B is used to compute the contact prediction matrix $\hat{C} \in [0,1]^{n \times m}$, where

$$\widehat{C}_{i,j} = \sigma(\text{Batch}(\text{Conv}(B_{(i-w:i+w),(j-w:j+w)}))) \quad \text{(Equation 5)}$$

The two-dimensional convolution (Conv) operation with width $2w+1$ and h channels uses the h-dimensional representation of all residues within w of $B_{i,j}$ to compute $\widehat{C}_{i,j}$, and thus detects local patterns in two-dimensional residue contact space. The broadcast matrix is zero-padded to allow for the convolution operation to be performed at all indices. We again apply a batch normalization to stabilize learning. We apply the sigmoid operation σ, which restricts the output values of $\widehat{C}$ to be in the range [0,1], and thus they can be interpreted as the predicted probability that two residues are in contact.

Interaction prediction module

The interaction prediction module computes a single probability of interaction $\hat{p}$ from the $n \times m$ contact map $\widehat{C}$. To do so, we perform two pooling operations. The first is a standard max-pool: an l-dimensional max-pool divides $\widehat{C}$ into $\left\lceil \frac{n}{l} \right\rceil \times \left\lceil \frac{m}{l} \right\rceil$ non-overlapping regions and takes the maximum value of each region, with zero-padding applied where necessary. This max-pooled matrix P represents the probability of interaction in local regions of the proteins and maintains only the highest-probability residue contacts in each region for global prediction. The second pooling operation is a global pooling operation, calculated as

$$Q_{i,j} = \text{ReLU}\left(P_{i,j} - \mu - \left(\gamma * \sigma^2\right)\right) \quad \text{(Equation 6)}$$

$$\hat{p}_{raw} = \frac{\sum_{i,j} Q_{i,j}}{\sum_{i,j}(\text{sign}(Q_{i,j})) + 1} \quad \text{(Equation 7)}$$

where μ, σ^2 are the mean and variance of the P_{ij} values and γ is a learned parameter. The matrix Q sparsifies P, maintaining only those contacts which are $\gamma\sigma^2$ greater than the mean, and setting all others to zero. We then predict that the proteins will interact with the average probability of interaction among these high-probability contacts. Together with the regularization that the contact matrix be sparse, this global pooling operation captures the intuition that a pair of interacting proteins will be characterized by a relatively small number of high-probability interacting residues or regions.

The final step of interaction prediction is designed to enhance the bimodality of the output distribution, so that the choice of a cutoff becomes less important in distinguishing positive and negative predictions. We apply the logistic activation function to compute the output probability $\hat{p} = \sigma_{(x_0,\eta)}(\hat{p}_{raw})$ where

$$\sigma_{(x_0,\eta)}(x) = \frac{1}{1 + e^{-\eta(x - x_0)}} \quad \text{(Equation 8)}$$

This activation function, with $x_0 = 0.5$, takes our raw probability predictions and makes them more "extreme", depressing values below x_0 towards 0 and inflating values above x_0 towards 1, with η controlling the rate at which this occurs. We return $\hat{p}$ and $\widehat{C}$ as the model prediction, from which we calculate the loss and optimize the gradient.

Training

Training objective

Given the true labels, the predicted probabilities $\hat{p}$, and the contact maps $\widehat{C}$, we compute the loss as $\lambda L^{BCE} + (1 - \lambda)L^{MAG}$; here λ is a hyper-parameter that balances between L^{BCE}, the binary cross-entropy (BCE) loss, and L^{MAG}, the contact-map magnitude loss (MAG). While the BCE loss is standard in a classification context, we introduce L^{MAG} as a regularization term that enables us to learn realistically sparse contact maps. L^{MAG} for a single training example is calculated as the arithmetic mean value of the contact map $\widehat{C}$. Jointly minimizing the total magnitude of contact maps with the BCE captures the intuition that interacting proteins are characterized by just a few high probability inter-protein contacts, while most residues will not be in contact. There has been substantial work on predicting single-protein, and, in some cases, protein-complex contact maps. Future work could explore incorporating some of these approaches into D-SCRIPT. Single-protein contact maps should be reasonably amenable to the D-SCRIPT model: one would simply augment the featurization to include the per-residue contacts. However, integrating pre-built protein-complex contact maps is trickier. The contact maps learned by D-SCRIPT not only capture *where* an interaction might happen, but also *if* an interaction might happen. We induce this behavior by our regularization term that incentivizes an empty contact map, all else being equal. This important aspect would be need to be accounted for (possibly with an appropriate transfer learning or regularization step) when using off-the-shelf contact-map predictions as inputs.

Implementation details

We implemented D-SCRIPT in PyTorch 1.2.0 and trained with a NVIDIA Tesla V100 with 32GB of memory. Embeddings from the pre-trained Bepler and Berger model were produced by concatenating the final values of the output and all hidden layers, so that $d_0 = 6165$. We used a projection dimension of $d = 100$, a hidden dimension of $h = 50$, a convolutional filter with width $2w + 1 = 7$, and a local max-pooling width of $l = 9$. We used $x_0 = 0.5, \eta = 20$ for the custom logistic activation, and $\lambda = 0.35$ for calculating the training loss. Weights were initialized using PyTorch defaults. We used a batch size of 25, the Adam optimizer with a learning rate of 0.001, and trained all models for 10 epochs.

Comparison of model complexity

The version of PIPR that we compare to has 72,500 trainable parameters. The full D-SCRIPT model has 629,207 trainable parameters, but the vast majority of those (616,600) are in the projection module, which is a simple linear combination of all concatenated

hidden states of the Bepler & Berger language model (which are themselves already redundant). The remaining stages of the model together have 12,707 trainable parameters. Practically, we prevent over-fitting through a high rate of dropout (50%) in the Projection module, combined with a very simple but structurally informed architecture in the Contact and Interaction prediction modules. Additionally, the contact map magnitude loss acts as a regularization by requiring that the predicted contact maps be sparse. Empirically, D-SCRIPT generalizes much better than PIPR, and does not seem to overfit to the training data.

Analysis

Assessing functional module coherence

Detection of protein functional modules was performed by spectral clustering, with pairwise distances between proteins assessed using Cao et al.'s diffusion state distance (DSD) metric (Cao et al., 2013, 2014). We generated 500 clusters, removed clusters with fewer than 3 nodes, and recursively split clusters with greater than 100 nodes. This module detection approach performed well in a recent DREAM challenge on functional module detection (Choobdar et al., 2019). Proteins were annotated with functions using Gene Ontology (GO) terms from FlyBase (Thurmond et al., 2019), filtering out electronically-inferred and homology-based annotations. All GO terms were mapped to a limited set of GO Slim terms using the *D. melanogaster* species-specific list (Adams et al., 2000). For each cluster, we computed the within-cluster functional similarity, calculated as the mean Jaccard similarity of the sets of GO Slim annotations for all pairs of proteins within the cluster. We also used a different, graph-theoretic measure of protein similarity based on GO terms (Zhao and Wang, 2018); it produced similar results (Figure S2). The distribution of within-cluster similarity scores were compared using a one-tailed t-test, with the null hypothesis that the 384 modules in the D-SCRIPT network and 374 modules in the PIPR network have the same average within-cluster similarity.

Predicting PPIs in cow

We selected 24,195 bovine proteins by selecting the longest isoform of each gene and, using HMMER (Eddy, 2009) and GODomainMiner (Alborzi *et al.*, 2018), limited ourselves to proteins that had at least one Pfam (Sonnhammer *et al.*, 1997; Finn et al., 2014) domain with some associated GO annotation. In general, this filtering step is unnecessary but here it allowed us to focus our computational resources on proteins likely to be interest. Of the set of 292.7 million possible pairwise combinations, we generated a list of fifty million candidate interactions by randomly sampling protein pairs; we included a special check to ensure that the 12 pre-identified rumen related genes were fully covered. We predicted the probability of interaction using the human-trained D-SCRIPT model, and selected edges with a predicted value greater than or equal to 0.5 as positive edges. This resulted in a predicted network of 476,399 positive interactions between 17,811 proteins with an average node degree of 53.5. We created functional protein modules by performing spectral clustering on the proteins, using the diffusion state distance (DSD, Cao et al., 2013) metric. For each module, we applied gene set enrichment analysis using the g:GOSt tool on the g:Profiler web server (Raudvere et al., 2019) to identify functions over-represented among the proteins in each cluster.

We analysed bovine gene expression using bulk RNA-seq data from 93 tissue-specific bovine samples (Halstead et al. 2021; GEO Accession GSE160028) and normalized each sample to counts per million (CPM). For each cluster, we identified genes which are expressed more highly in rumen tissue than on average across all tissues, and compute and report the log fold increase in expression in the rumen. The gene expression data was also used to verify the quality of our predicted network and modules. We computed the Spearman rank correlation between the CPM normalized expression profiles for each pair of genes, then evaluated the average correlation between protein pairs with no interaction predicted, pairs with a predicted positive edge, pairs which are predicted to interact and share a cluster, and pairs of proteins which appear in the same module, regardless of whether or not an edge appears. We perform a one-sided Welch's t-test to compare sample means without assuming equal variance.

Identifying the catalytic signature motif of bovine protein tyrosine phosphates from D-SCRIPT predictions

Our analysis of functional modules in the bovine rumen found a single cluster containing 34 proteins homologous to human protein tyrosine phosphatases (PTPs). PTPs are known to comprise several families with similar sequence but diverse binding specificity, determined in part by a short catalytic signature motif (Alonso et al., 2004; Alonso and Pulido, 2016). In Figure S4, we show Cluster A from Figure 4A, recolored based on the PTP sub-type. All but one of the neighbors of *PRD-SPRRII* in Cluster A are PTP proteins (the exception being *MARK2*, a serine/threonine-protein kinase). We find that D-SCRIPT does not limit its predictions to one family, instead predicting interactions in all sub-types.

To further investigate how D-SCRIPT determines binding specificity, we undertook an *in silico* mutagenesis experiment. The canonical catalytic signature motif for the PTP family is HCX_5R (Kim and Ryu, 2012) or HCXXGXXR (Alonso and Pulido, 2016), a motif which we identified in 28 PTP proteins from Cluster A. We also included ENSBTAP00000067545 (*CDC25C*), which has the motif HCX_5A in our analysis. For each protein, we used D-SCRIPT to predict the probability of interaction with ENSBTAP00000070493 (*PRD-SPRRII*). Then for 50 replicates, we randomly perturbed the catalytic motif in that protein, either by randomly selecting amino acids for all 8 positions of the motif, or only the 5 flexible positions (X). We find that for 24 of the 29 proteins, perturbing only the flexible positions does not reduce the D-SCRIPT predicted probability, while perturbing the entire motif drastically decreases the predicted probability. For 2 of the remaining 5, D-SCRIPT already did not predict an interaction with the original protein, for another one perturbing even the flexible positions decreased the probability of interaction substantially, and for the final 2 even perturbing the entire motif did not substantially decrease the predicted probability of interaction. Figure S5A shows the original prediction (red), the distribution of 50 replicates where only flexible sites were mutated (blue), and the distribution of 50 replicates where the entire catalytic motif was mutated (orange) for each PTP protein.

Finally, we sought to identify which residues were most important in determining the D-SCRIPT model's prediction. To do so, we selected 5 CDC14 proteins (ENSBTAP00000058782, ENSBTAP00000073534, ENSBTAP00000054725, ENSBTAP00000069880, ENSBTAP00000070948) and aligned them using MUSCLE (Edgar, 2004). Then, for each position in the alignment, for all sequences which didn't have a gap in that position, we randomly perturbed the amino acid at that position 50 times and used D-SCRIPT to predict interaction between the perturbed sequence and *PRD-SPRRII*. For each sequence location, we computed the difference between the original predicted probability of interaction and the average probability of interaction predicted for the samples perturbed at that location, averaged across all 5 sequences (Figure S5B). We found the largest decreases around the catalytic signature motif, indicating that D-SCRIPT is in fact basing its predictions on the residues involved in binding specificity. Further, when we zoom in to the 8-residue motif region, it is clear that D-SCRIPT is identifying the conserved "C" and "R" in the second and eighth position, and the flexible "A" in the fourth position of the motif as the most important residues for determining interaction. Figure S5C shows the WebLogo for this region in the PTP domain, where the y-axis is the change in predicted probability when each position is perturbed (Schneider and Stephens, 1990; Crooks et al., 2004).

Predicting PPIs in SARS-CoV-2

We performed a preliminary study to predict viral-host interactions between SARS-CoV-2 and human proteins wherein we compared the sets of over-represented GO terms for human interactors of SARS-CoV-2 proteins, as predicted by D-SCRIPT or PIPR, with those over-represented in the experimentally-determined human interactors (Gordon et al., 2020). Figure S3 shows the relative similarity of computationally predicted annotations to the experimentally-determined annotations for each SARS-CoV-2 protein. Overall, we found that sets of enriched terms computed using the D-SCRIPT network overlap slightly more with the true network than those computed using the PIPR network ($p = 0.059$). Among the putative accessory factors (ORF* and Protein 14), D-SCRIPT performs significantly better (mean Jaccard similarity 0.029 vs. 0.118, $p = 0.022$, paired one-tailed t-test). Visually, PIPR seems to be somewhat better at predicting interaction partners for the non-structural proteins (NSP*), although D-SCRIPT still has a slightly larger mean similarity (0.183 vs. 0.222, $p = 0.221$). While D-SCRIPT performs better on the intensively studied spike (S) protein, PIPR shows a higher overlap for the nucleocapsid (N). Neither method predicts enriched terms for the other structural proteins encoding the envelope (E) and membrane (M) (0.149 vs. 0.121, $p = 0.672$ across the four proteins).

Methods

Candidate pairs were generated using the viral sequences from Gordon et al. and 19,777 human sequences from the STRING database, and predicted edges using D-SCRIPT and PIPR. We predicted 3,273 edges using D-SCRIPT and 2,922 edges using PIPR. 332 putative true viral-host interactions were taken from Gordon et al. Human sequences were mapped to UniProt sequences identifiers from Gordon et al. with sequence similarity $\geq 95\%$ using BLAST (Altschul et al., 1990), and UniProt identifiers were used to identify a set of Gene Ontology terms for the human interactors of each viral protein. Following Gordon et al., we identified over-represented GO terms using the clusterProfiler R package (version 3.14.3) (Yu et al., 2012) with a 1% false discovery rate (FDR). Over-represented GO terms were mapped to a common set of terms taken from the ChEMBL Drug Target GO Slim Subset (Mutowo et al., 2016). For each viral protein, we computed the Jaccard similarity between the set of GO Slim terms enriched in the putative true network and each of the computationally predicted methods. We computed a paired one-tailed t-test to statistically compare the relative similarities of D-SCRIPT and PIPR. Virus-host edges predicted using D-SCRIPT or PIPR are available for use by the community (key resources table).

Logistic regression for prediction of protein structure

We selected 300 proteins with structural coordinates from the Protein Data Bank (PDB), and randomly split them into a training set of 100 and test set of 200 proteins. We assessed intra-protein contacts at 8 Å in the PDB structure and converted each protein's contact map to a binary classification data set: a protein sequence of length n corresponded to $\frac{n(n-1)}{2}$ observations, with the observation ij corresponding to a putative contact between residues i and j. Features were generated using the human pre-trained D-SCRIPT model, where we evaluate the model up through the first stage (projection module). This generates $d = 100$ dimensional vector representations Z_i and Z_j output by the projection module for each pair of amino acids in the protein which capture both local and global structural features. The regression was L_2-regularized and class balanced, with its input for observation ij being the concatenation of the 100-dimensional embeddings as well as their combinations diff_{ij}, mul_{ij} as defined in Equations 2 and 3.

Predicting interactions by nearest neighbor search

To compare our sequence embedding method with other potential ones, we evaluated various protein sequence embeddings under the following framework: the Euclidean distance in an embedding space was used to define a distance measure between proteins. Given a true positive PPI (A,B), we applied this distance measure to identify the k-nearest neighbors of A and B each, and computed how many of the k^2 possible combinations of these neighbors corresponded to a positive PPI. For D-SCRIPT, we used the $\mathcal{R}_{\geq 0}^{n \times d}$ output of the projection module and averaged the features across the length of the protein to obtain a d-dimensional embedding. For AAClass, Vec5, and the random embedding, we directly compute the Euclidean distance between those vectors. To identify k nearest neighbors using BLAST, we create a database of all other proteins in the species, perform a search using blastp with the default values, and return the top k hits by e-value.

Predicting inter-protein contact maps

The output of the residue Contact Module of D-SCRIPT is an inter-protein contact map $\widehat{C}$ where $\widehat{C}_{ij} \in [0, 1]$ can be interpreted as the probability of residue i from protein S_1 being in contact with residue j of protein S_2. We interpreted ground truth and predicted contact maps as probability distributions over the $n \times m$ matrix and measured the 2-D Earth mover's distance between these distributions, computed by solving an optimal transport problem under the Euclidean metric (Flamary and Courty, 2017). For each candidate

PPI, we established random baselines by shuffling $\hat{C}$ 500 times and recomputing the Earth mover's distance between the random shuffle and the true contacts. We assign a p-value to each candidate PPI based on this permutation test and the probability of seeing an Earth mover's distance at least as small as the observed distance. We compute an overall p-value for positive-predicted pairs by computing a one-tailed t-test with the null hypothesis that the average candidate PPI p-value is 0.5, i.e. that the D-SCRIPT predictions are as accurate as a random shuffling.

Deconstructing model performance by protein frequency in training data

To further investigate the relative performance of each model on out-of-species classification, we evaluated each model on subsets of proteins ranked by their frequency in the training set. For a set of quantiles $q \in \{0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$, we evaluated out-of-sample D-SCRIPT and PIPR predictions on the human PPI sub-network consisting only of proteins of quantile q or lower; here, lower q corresponds to a lower frequency of occurrence. In absolute terms, as Table S2 indicates, both D-SCRIPT and PIPR become more accurate at higher q. However, D-SCRIPT has a relative advantage at lower q (i.e., infrequently occuring proteins) while PIPR performs better at higher q. In other words, PIPR's better within-species performance can be traced to it being more accurate on proteins that occur frequently. This also suggests an explanation for PIPR's lower cross-species generalizability than D-SCRIPT: when making predictions on an entirely new set of proteins in a different species, knowing the relative frequencies of proteins in the training data might not be particularly useful.

The difference between D-SCRIPT and PIPR might stem from their respective architectures. The protein representation learned by D-SCRIPT is constrained to be a linear projection of the Bepler & Berger pre-trained embedding, albeit with ReLU and dropout layers. This regularizes how much frequency information can be incorporated into the model; we note that the Bepler and Berger model was trained with data on individual proteins and would not reflect PPI frequency information. In contrast, PIPR's design allows for a lot more leeway in training each protein's representation. This flexibility may allow PIPR to better incorporate the occurrence frequencies into its representation, helping its within-species performance but potentially hurting its cross-species generalizability.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical tests were conducted using version 1.3.1 of the SciPy Python package.