Deep Generative Models of Protein Structure Uncover Distant Relationships Across a Continuous Fold Space

Eli J. Draizen[1,2*], Stella Veretnik[1], Cameron Mura[1,2*], Philip E. Bourne[1,2]

[1]School of Data Science, University of Virginia, Charlottesville, VA, USA.
[2]Department of Biomedical Engineering, University of Virginia, Charlottesville, VA, USA.

*Corresponding author(s). E-mail(s): edraizen@gmail.com; cmura@virginia.edu;

Abstract

Our views of fold space implicitly rest upon many assumptions that impact how we analyze, interpret and understand biological systems—from protein structure comparison and classification to function prediction and evolutionary analyses. For instance, is there an optimal granularity at which to view protein structural similarities (e.g., architecture, topology or some other level)? If so, how does it vary with the type of question being asked? Similarly, the discrete/continuous dichotomy of fold space is central in structural bioinformatics, but remains unresolved. Discrete views of fold space bin 'similar' folds into distinct, non-overlapping groups; unfortunately, such binning may inherently miss many remote relationships. While hierarchical databases like CATH, SCOP and ECOD represent major steps forward in protein classification, a scalable, objective and conceptually flexible method, with less reliance on assumptions and heuristics, could enable a more systematic and nuanced exploration of fold space, particularly as regards evolutionarily-distant relationships. Building upon a recent 'Urfold' model of protein structure, we have developed a new approach to analyze protein interrelationships. Termed 'DeepUrfold', this method is rooted in deep generative modeling via variational Bayesian inference, and we find it to be useful for comparative analysis across the protein universe. Critically, DeepUrfold leverages its deep generative model's learned embeddings, which occupy high-dimensional latent spaces and can be distilled for a given protein in terms of an amalgamated representation that unites sequence, structure, biophysical and phylogenetic properties. Notably, DeepUrfold is structure-*guided*, versus being purely structure-based, and its architecture allows each trained model to learn protein features (structural and otherwise) that, in a sense, 'define' different superfamilies. Deploying DeepUrfold with CATH suggests a new, mostly-continuous view of fold space—a view that extends beyond simple 3D structural/geometric similarity, towards the realm of integrated *sequence*↔*structure*↔*function* properties. We find that such an approach can quantitatively represent and detect evolutionarily-remote relationships that evade existing methods.

Availability: Our results can be explored in detail at https://bournelab.org/research/DeepUrfold. The DeepUrfold code is available at http://www.github.com/bouralab/DeepUrfold, and associated data are available at https://doi.org/10.5281/zenodo.6916524.

Keywords: deep learning; evolution; fold space; generative model; protein structure; protein classification; remote homology

Introduction

The precise historical trajectory of the protein universe [1] remains quite murky, and likely corresponds to an evolution from (proto-)peptides, to protein domains, to multi-domain proteins [2]. Presumably, the protein universe—by which we mean the set of all unique protein sequences (known or unknown, natural or engineered, ancestral or extant)—did not spontaneously arise with intact, full-sized domains. Rather, smaller, sub-domain–sized protein fragments likely preceded more modern domains; the genomic elements encoding these primitive fragments were subject to natural evolutionary processes of duplication, mutation and recombination to give rise to extant domains found in contemporary proteins [2–6]. Our ability to detect common polypeptide fragments, shared amongst at least two domains (in terms of either sequence or structure), relies upon having (i) a similarity metric that is sensitive and accurate, and (ii) a suitable random/background distribution (i.e., null model) for distances under this metric; historically, such metrics have been rooted in the comparison of either amino acid sequences or three-dimensional (3D) structures, often for purposes of exploring protein fold space. The recent advent of high-accuracy structure prediction [7, 8], enabled by deep learning, presents new opportunities to explore fold space; to do so effectively requires new methods to accurately and sensitively detect weak/distant relationships.

Fold Space, Structural Transitions & Protein Fragments

Fold space[1], as the collection of all unique protein folds, corresponds to a many-to-one mapping: vast swaths of sequence space map to fold $\mathcal{A}$, another vast swath maps to fold $\mathcal{B}$, a narrower range might map to fold $\mathcal{C}$, and so on. Two proteins that are

[1] The term "protein structure space" (PSS) means the set of all protein 3D structures, known and unknown; the term "fold space" refers to the set of all protein folds. Though not strictly equivalent [12], we treat these terms interchangeably here unless noted otherwise.

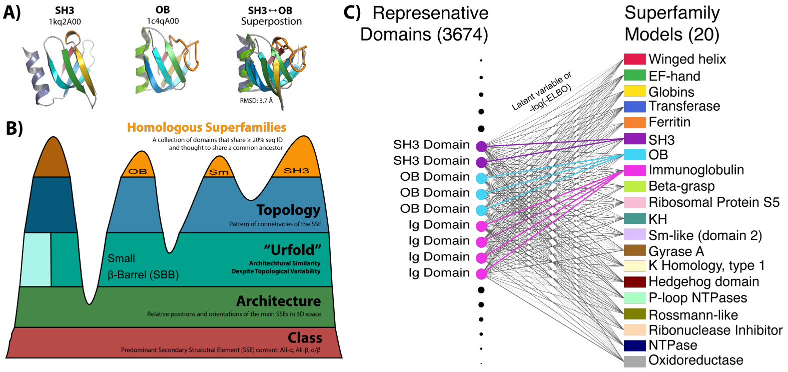

Fig. 1 Overview of the Urfold model and DeepUrfold approach to identify domains that might reflect the phenomenon of "*architectural similarity despite topological variability*". (A) The SH3 and OB domains are prototypical members of the small β-barrel (SBB) urfold because they have the same barrel architecture, yet different strand topologies: they have strikingly similar 3D structures and share extensive functional similarities (e.g., PPI binding on the same edge-strand, involvement in nucleic acid–binding and processing pathways [9, 10]), yet these similarities are obscured by the SH3 and OB superfolds having been classified differently. In the case of the SBB urfold, the loops linking the strands are permuted in the SH3 and OB, yielding the different topologies seen in their 3D superposition. (B) If the Urfold phenomenon is viewed in terms of CATH, it is hypothesized to be a discrete structural entity or 'level' that lies between the Architecture and Topology strata, as schematized here. (C) DeepUrfold, which applies deep learning to the Urfold conceptualization of protein structure, identifies new potential urfolds by creating 20 SF-specific VAE neural network models and comparing output scores from all representative domains from those superfamilies (numbering 3,674) to every other SF model. As a metric to compute initially, we can imagine comparing the latent variables from domain representatives using models trained on the same SF (colored lines; see Fig. 3). Then, we can perform an all-vs-all comparison to begin mapping fold space, which we view as being organized as mixed-membership communities, versus hierarchically-clustered, mutually-exclusive bins; as detailed below and illustrated in Fig. 4, such communities can be computed via stochastic block models (SBMs; reviewed in [11]).

closely related (evolutionarily) might adopt quite similar folds ($\mathcal{A}$, $\mathcal{B}$), leading to their proximity in this high-dimensional space. Traditionally, fold space has been examined by hierarchically clustering domains based upon 3D structure comparison; in such approaches, whatever metric is used for the comparison can be viewed as structuring the space. The transition of a protein sequence from one fold to another, whether it be nearby ($\mathcal{A} \rightarrow \mathcal{B}$) or more distant ($\mathcal{A} \rightarrow \mathcal{C}$), and be it naturally (via evolution) or artificially (via design/engineering), likely occurs over multiple intermediate steps. These mechanistic steps include processes such as combining or permuting short secondary structural segments or longer regions (such as whole secondary structural elements [SSEs]), or mutating individual residues via nonsynonymous substitutions [5, 13–16]. In general, each such step may yield a new 3D structure, and that structure may correspond to the same or a different 'fold'. Similarities across these transitional states, $\mathcal{A} \rightarrow \mathcal{A}' \rightarrow \mathcal{A}'' \rightarrow \cdots \rightarrow \mathcal{B}$, blur the boundaries that delineate distinct groups—increasing or decreasing a relatively arbitrary and heuristic quantity, such as an RMSD or other similarity threshold, effectively alters the granularity of groupings in this space, and can change which structures belong to which groups. In this sense, the discrete versus continuous duality of fold space can be viewed largely as a matter of semantics or thresholding, versus any 'real' (intrinsic or fundamental) feature of the space itself [17].

Despite their limitations, it was pairwise similarity metrics in structure space that first indicated remote connections in a continuous fold space via shared fragments (see [18] and references therein). In an early landmark study, Holm & Sander [19] created an all-by-all similarity matrix from 3D structural alignments and discovered that the protein universe harbors five peptide 'attractors', representing frequently-adopted folding motifs (e.g., the β-meander). Nearly a decade later, and armed with vastly more 3D structures, similar pairwise analyses across protein structure space showed that 'all-α' and 'all-β' proteins are separated by 'α/β' proteins [20]. All-by-all similarity metrics applied to full domains (or fragments thereof) can be equivalently viewed as a graph-theoretic adjacency matrix, thus enabling the creation of a network representation of fold space. Such networks have been found to be "nearly connected", linking various domains (graph nodes) in ≈4-8 hops [21–23].

Graph-based representations of individual proteins have also motivated the study of common short (sub-domain) fragments. In pioneering studies, Harrison et al. [24, 25] found maximal common cliques of connected SSEs in a graph-based protein representation; their model took SSEs (helices, strands) as vertices and used the pairwise geometric relationships between SSEs (distances, angles, etc.) to decorate the graph's edges. In that work, 80% of folds were found to share common cliques with other folds, and these were quantified by a new concept termed '*gregariousness*'.

Although short, sub-domain–sized peptide fragments have been thoroughly studied, relatively few approaches have taken an evolutionary perspective, in the context of a continuous fold space. Goncearenco et al. [26] identified common loop fragments flanked by SSEs, called '*elementary functional loops*' (EFLs), that couple in 3D space to perform enzymatic activity. Youkharibache [6] noticed that peptide fragments, called '*protodomains*', are often composed (with C_2 internal symmetry) to give a larger, full-sized domain. More recently, Bromberg et al. identified common fragments between metal-binding proteins using '*sahle*', a new length-dependent structural alignment similarity metric [4]. These studies underscore the functional (and thus evolutionary) roles of sub-domain structural fragments.

The two state-of-the-art, evolution-based fragment libraries that are currently available, namely '*primordial peptides*' [2] and '*themes*' [27], involved creation of a set of common short peptide fragments based on HHsearch [28] profiles for proteins in SCOP and ECOD, respectively. The sizes of the libraries created by these two sequence-driven approaches (40 primordial peptides, 2195 themes) vary greatly, reflecting different stringencies of thresholds (and, ultimately, their different goals).

Another approach to study shared, commonly-occurring sub-domain fragments is to represent a protein domain as a vector of fragments. For example, the *FragBag* method [29] describes a protein by the occurrence of fragments in a clustered fragment library [30]. A recent and rather unique approach, *Geometricus* [31], creates protein embeddings by taking two parallel approaches to fragmentation: (i) a k-mer based fragmentation runs along the sequence (yielding contiguous segments), while (ii) a radius-based fragmentation uses the method of spatial moment invariants to compute (potentially non-contiguous) geometric 'fragments' for each residue position and its neighborhood within a given radius, which are then mapped to 'shape-mers'. Conceptually, this allowance for discontinuous fragments is a key step in allowing an algorithm to bridge more of fold space, as similarities between such non-contiguous fragments can imply an ancestral (contiguous) polypeptide that duplicated and lost one or more N'- or C'-terminal SSEs, perhaps in a "creative destruction" process that yields two different folds (i.e., different topologies) despite the preserval of similar architectures [5, 16].

Limitations of Hierarchical Systems, and the Urfold

The conventional view of fold space as the constellation of all protein folds, grouped by their 'similarities' to one another, largely rests upon hierarchically clustering domains based upon 3D structure comparison, as exemplified in pioneering databases such as CATH [32], SCOP [33, 34], and ECOD [35]. Despite being some of the most comprehensive and useful resources available in protein science, these databases have intrinsic limitations that stem from their fundamental structuring scheme, reflecting assumptions and constraints of any hierarchical system (e.g., assigning a given protein sequence to one mutually exclusive bin versus others); in this design schema, domains with the same fold or superfamily (SF) cluster discretely into their own independent 'islands'. The difficulty in smoothly traversing fold space, at least as it is construed by these databases—e.g., hop from island-to-island or create 'bridges' between islands in fold space—implies that some folds have no well-defined or discernible relationships to others. That is, we miss the weak or more indeterminate (but nevertheless *bona fide*) signals of remote relationships that link distantly-related folds. In addition to the constraints imposed by mutually exclusive clustering, the 3D structural comparisons used in building these databases generally rely upon fairly rigid spatial criteria, such as requiring identical topologies for two entities to group together at the finer (more homologous) classification levels. *What relationships might be detectable if we relax the constraints of strict topological identity?* As described below, this question is addressed by a recently proposed 'Urfold' model of protein structure [9, 12], which allows for sub–domain-level similarity.

Motivated by sets of striking structure↔function similarities across disparate superfamilies, we recently identified relationships between several SFs that exhibit architectural similarity despite topological variability, in a new level of structural granularity that allows for discontinuous fragments and that we termed the 'Urfold' (Fig. 1B; [9, 12]). Urfolds[2] were first described in the context of small β-barrel (SBB) domains (Fig. 1A), based on patterns of structure↔function similarity (as well as sequence signatures in MSAs, albeit more weakly) in deeply-divergent collections of proteins that adopt either the SH3/Sm or OB superfolds [9]. Notably, the SH3 and OB are two of the most ancient protein folds, and their antiquity is reflected in the fact that they permeate much of information storage and processing pathways (i.e., the transcription and translation apparatus) throughout all three domains of cellular life [16, 36, 37].

[2] We use the capitalized term 'Urfold' to refer to the concept/theory/model, as a general idea; the lowercase 'urfold' is used when we intend for that specific instance of the word to be limited to a specific case (e.g., "*the* SBB urfold"). Our goal is not to be dogmatic, but rather to be clear and precise as this new concept is being developed.

DeepUrfold: Motivation & Overview

The advent of deep learning [38], including the application of such approaches to protein sequences and structural representations, affords opportunities to study protein interrelationships in a wholly new and different way—namely, via quantitative comparison of 'latent space' representations of a protein in terms of its lower-dimensional 'embedding'. Such embeddings can be derived at arbitrary levels of granularity (e.g., atomic) and can subsume virtually any types of properties, such as amino acid type, physicochemical features (e.g., electronegatitivty), geometric attributes (e.g., surface curvature), phylogenetic conservation of sites, and so on. Two powerful benefits of such approaches are that (i) models can be formulated and developed in a statistically well-principled manner (or at least strive to be clear about their assumptions), and (ii) models have the capacity to be *integrative*, by virtue of the encoding (or 'featurization') of structural properties alongside phylogenetic, chemical, etc. characteristics of the data (in this case, augmenting purely 3D structural information about a protein). The methodology presented here explores the idea that viewing protein fold space in terms of feature embeddings and latent spaces (what regions are populated, with what densities, etc.)—and performing comparative analysis via such spaces (versus in direct or 'real' 3D/geometric space)—is likely to implicitly harbor deep information about protein interrelationships, over a vast multitude of protein evolutionary timescales. Such distant timescales are likely to be operative at the Urfold level of structure [12].

Here, we present a deep learning–based framework, 'DeepUrfold', to systematically identify urfolds by using a new alignment-free, topology-agnostic, biochemically-aware similarity metric of domain structures, based on deep generative models, together with mixed-membership community detection algorithms. From a probabilistic perspective, our metric is rooted in the variational Bayesian inference that underpins variational autoencoders (VAEs [39]). From a deep learning perspective, our algorithm leverages embeddings and similarities in latent-space representations rather than simple (purely-geometric) 3D structures directly, enabling us to encode any sort of biophysical or other types of properties and thereby allowing more subtle patterns of similarities to be detected—such as may correspond to architectural similarities among (dis-)contiguous fragments from different folds, or even superfolds, that are related only at great evolutionary distances (Fig. 1C).

In brief, DeepUrfold's four distinct methodological stages are: (i) *Dataset construction*, whereby 3D structures are prepared, featurized and allocated into suitable training/test splits for machine learning; (ii) *Training of SF-specific models*, using featurized protein structural data and a hybrid 3D-CNN/VAE-based deep network; (iii) *All-by-all inference calculations*, computing VAE-derived ELBO-based scores to assess the 'fit' of each 3D structural domain representative to each SF (i.e., subject each SF representative, i, to each SF-specific VAE model, j); (iv) *Elucidation of any community structure* in these protein $\rightsquigarrow$ SF mappings, via stochastic block modelling of the patterns of scores.

Results

The DeepUrfold Computational Framework: Deep Generative Models

Conventionally, two protein structures that have similar architectures but varying topologies (i.e., folds) might be viewed as having resulted from convergent evolution. However, as in the case with the SH3 and OB superfolds, the *structure* ↔ *function* similarities [9], and even *sequence* ↔ *structure* ↔ *function* similarities [16], can prove to be quite striking, suggesting that these domain architectures did not arise independently [6, 16] but rather are echoes of a (deep) homology. To probe what may be even quite weak 3D similarities, in DeepUrfold we model the evolutionary processes giving rise to proteins as an integrated 3D structure/properties 'generator'. In so doing, we seek to learn probability distributions, $p(x|\theta)$, that describe the specific geometries and physicochemical properties of different folds (i.e., features that largely define protein *function*), where the random variable x denotes a single structure drawn from ($x \in \mathbf{x}$) a set of structures labelled as having the same fold ($\mathbf{x}$), and θ denotes the collection of model parameters describing the variational distribution over the background (i.e., latent) parameters. We posit that folds with similar latent space embeddings and learned probabilistic distributions—which can be loosely construed as "structure ↔ function mappings", under our feature-set—likely have similar geometries/architectures and biophysical properties, regardless of potentially differing topologies (i.e., they comprise an urfold), and that, in turn, may imply a common evolutionary history.

Using the principles of variational inference [40], DeepUrfold learns the background distribution parameters $\boldsymbol{\theta_i}$ for superfamily distributions, i.e. models $p_i(x_{ij}|\boldsymbol{\theta_i})$, by constructing and training a variational autoencoder (VAE) model for each superfamily i and domain structure j. In the current work, DeepUrfold is developed using 20 highly-populated SFs from CATH (see Fig 1C and Supp Table 1). The original/underlying likelihood distribution, $p_i(x_{ij}|\boldsymbol{\theta_i})$, is unknown and intractable, but it can be estimated by considering an easier-to-approximate posterior distribution of latent space parameters, $q_i(z_{ij}|\mathbf{x}_i)$, where z denotes the latent variables we wish to infer and, again, $\mathbf{x}$ is our data (protein structures); in our case, the approximating distribution $q(z|\mathbf{x})$ is taken as sampling from a Gaussian. To ensure that $q_i(z_{ij}|\mathbf{x}_i)$ optimally describes $p_i(x_{ij}|\boldsymbol{\theta_i})$, one can seek to maximize an *evidence lower bound* (ELBO) quantity as a variational objective, which supplies a lower bound of the marginal log-likelihood of a single structure, $\ln[p_i(x_{ij})]$. The ELBO inequality can be written as:

$$\begin{aligned}\ln[p_i(x_{ij})] \geq\ & \mathbb{E}_{q_i(z_{ij}|\mathbf{x}_i)}[\ln\ p_i(x_{ij}|z_{ij})] - \\ & D_{\mathrm{KL}}[q_i(z_{ij}|x_{ij})\,||\,p(z_{ij})]\end{aligned} \tag{1}$$

where $p_i(x_{ij})$ is the likelihood, $\mathbb{E}$ is the expectation value of q in terms of p, and $D_{\mathrm{KL}}[q||p]$ is the Kullback-Leibler divergence, or relative entropy, between the two probability distributions q and p. In other words, maximizing the ELBO corresponds to maximizing the expected log-likelihood of our learned model and minimizing the entropy or 'distance' (D_{KL}) between (i) the true/exact underlying prior distribution of the data given a model, $p(x|\boldsymbol{\theta})$, and (ii) our learned/inferred approximation, as a posterior distribution of latent parameters given the data, $q(z|\mathbf{x})$. Pragmatically, DeepUrfold's variational objective is formulated as a minimization problem (Supp Info §3), so we compute –(ELBO) values.[3] In a similar vein, part of DeepUrfold's testing and development (detailed below) involved training "joint models" using a bag of SFs with intentionally *different* topologies, e.g., a mixed SH3 ∪ OB set, while accounting for the class imbalance [41, 42] that stems from there being vastly different numbers of available 3D structural data for different protein SFs (e.g., immunoglobulin [Ig] structures, which are disproportionately abundant). Further details of the multi-loop permutation analyses used in testing and developing DeepUrfold can be found in Supp Info §4.

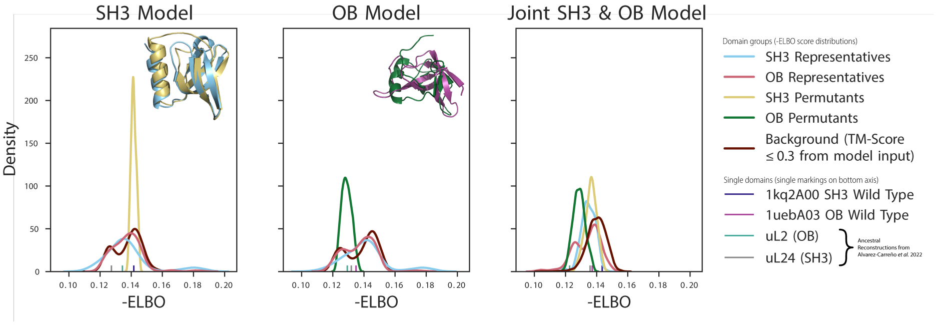

Fig. 2 Likelihood-based ELBO values can quantify similarities among multi-loop permuted structures. To gauge the sensitivity of DeepUrfold's VAE-based metric to loop orderings (topology), we generated a series of fictitious folds and analyzed their patterns of scores. Specifically, we implemented a multi-loop permutation algorithm [43] to systematically 'scramble' the SSEs found in an SH3 domain (1k2A00) and an OB domain (1uebA03); in these loop 'rewiring' calculations, we stitched together the SSEs and energetically relaxed the resultant 3D structures. While 96 unique permutations are theoretically possible for a 4-stranded β-sheet [9], only 55 SH3 and 274 OB permuted domains were able to be modeled, presumably because their geometries lie within the radius of convergence of MODELLER (e.g., the loop-creation algorithm did not have to span excessive distances in those cases). Each novel permuted structure was subjected to a DeepUrfold VAE model that had been trained on all other domains from either SH3-only (left panel), (B) OB-only (middle), or (C) joint SH3 ∪ OB domain (right) datasets. Fits to models were approximated by the –(ELBO) score, which can be viewed as a similarity metric or a measure of 'goodness-of-fit' between an individual structure and the SF-level VAE model trained via DeepUrfold. In reference to a given model, a given permutant query structure having a –ELBO less than its wild-type structure for that model can be considered as structurally more 'similar' (a better fit) to the model, and thus perhaps more thermodynamically or structurally stable. As reference points, we also include the –ELBO scores for ancestrally-reconstructed progenitors of the OB (uL2) and SH3 (uL24) superfolds, based on recent work by Alvarez-Carreño et al. [16]; note that these latter data are single 3D structures (not datasets/distributions of structures) subjected to a single inference pass through a trained VAE model, and therefore they appear as thin vertical 'tick' lines along the horizontal axis. See Supp Info §4.2 and Supp Fig S8 for further discussion of these traces, including interpretations of the background distributions (maroon traces) and the single-tick entities.

[3] This reasoning underlies the interpretation of Fig 2: Maximizing the ELBO equates to minimizing DeepUrfold's –(ELBO) loss function, which is why a shift leftwards along the horizontal axis in Fig 2 corresponds to 'better' models. Similarly, more positive values of the –(ELBO) quantity reflect poorer agreement between a domain structure and the VAE model it is being subjected to in an inference calculation (e.g., the single-tick marks in Fig 2); see also the analysis provided in Supp Info §4.2.

As input to the VAE, we encode the 3D structure of a protein domain by representing it as a 3D volumetric object, akin to the input used in 3D convolutional neural networks (CNNs). Indeed, DeepUrfold's neural network architecture can be viewed as a hybrid/stacked 3D CNN-based VAE. In our discretization, atoms are binned into volumetric elements (voxels), each of which can be tagged or labeled, atom-wise, with arbitrary properties (biophysical, phylogenetic, etc.). A critical point is that this representation scheme is agnostic of polypeptide chain topology, as the covalent bonding information between residues, and the order of SSEs, is not explicitly retained; note, however, that no information is lost by this representation, as such information is implicit in the proximity of atom-occupied voxels in a model (and can be used to unambiguously reconstruct a 3D structure). The above preparatory and featurization steps utilized 'Prop3D', a computational toolkit that we have developed for machine learning with protein structures [44].

Note that we do not use VAEs to generate new samples from a given SF per se. Rather, the role of the VAE in DeepUrfold can be viewed as that of an anomaly detection tool, to robustly and quantitatively address the question: "Based on learned, superfamily-specific latent space representations, what is the likelihood that a given domain structure (from any SF, i) arose from (or, alternatively, *was generated by*) a particular SF-specific VAE model, j?".

DeepUrfold Models Can Detect Similarities among Topologically-distinct, Architecturally-similar Proteins

To initially assess our SH3, OB and joint SH3/OB DeepUrfold models—and to examine the properties of the Urfold model more broadly—we directly probed the Urfold's core concept of "*architectural similarity despite topological variability*". This test was performed by considering sets of artificial protein domains that have identical architectures but with specifically introduced loop permutations; we obtained these systematically engineered perturbations of a 3D structure's topology by 'rewiring' the SSEs (scrambling the loops), while retaining the overall 3D structure/shape (i.e., architecture). Specifically, (i) we systematically created permuted (fictitious) 3D structures starting with representative SH3 and representative OB domains (Supp. Fig. 7A) via structural modeling (including energetic relaxation), and (ii) we then subjected each of these rewired structures, in turn, to each of the SH3, OB and joint SH3/OB DeepUrfold models. The SH3/Sm and OB superfolds comprise the first-identified urfold [9], namely the small β-barrel (SBB). While SBBs typically have six SSEs (five strands and a helix), there are four 'core' β-strands, meaning an SBB's β-sheet can theoretically adopt one of at least 96 distinct loop permutations [9]; note that, based on the operational definitions/usage of the terms 'topology' and 'fold' in systems such as SCOP, CATH, etc., such engineered permutants almost certainly would be annotated as being from different homologous superfamilies, implying no evolutionary relatedness. Thus, the loop-scrambling approach described here is a systematic way to gauge DeepUrfold's ability to discern similarities at the levels of architecture and topology, in a self-contained manner that is agnostic of preexisting classification schemes such as CATH.

In general, we find that the synthetic/permuted domain structures have similar distributions of −(ELBO) scores as the corresponding wild-type domains (Fig 2). Those permuted domain structures with −(ELBO) scores more negative than the wild-type domains (i.e., distributions that shift leftward in Fig 2 and Supp Fig S8) can be interpreted as being more similar (structurally, biophysically, etc.) to the DeepUrfold variational model (a 'consensus' model, of sorts[4]), and thus perhaps more thermodynamically stable or structurally robust were they to exist in reality—an interesting possibility as regards protein design and engineering. In terms of more conventional structural similarity metrics, the TM-scores [45] for permuted domain structures against the corresponding wild-type topolog (Supp Fig S7a) typically lied in the range $\approx 0.3-0.5$—i.e., values which would indicate that the permutants and wild-type are not from identical folds, yet are more than just randomly similar (Supp Fig S7b).

The findings from these test calculations suggest that the DeepUrfold model is well-suited to our task because our encoding is agnostic to topological connectivity information and, rather, is sensitive only to 3D spatial architecture/shape. Even though polypeptide connectivity is implicitly captured in our discretization, our DeepUrfold model intentionally does not consider if two residues are linked by a peptide bond or if two spatially proximal SSEs are contiguous in sequence. The generality of this approach is useful in finding similarities amongst sets of seemingly dissimilar 3D structures—and thereby identifying specific candidate urfolds—because two subdomain portions from otherwise rather (structurally) different domains may be quite similar to each other, even if the domains which they are a part of have different (domain-level) topologies but identical overall architectures. This concept can be represented symbolically: for an arbitrary subset of SSEs, d, drawn from a full domain $\mathcal{D}$, the Urfold model permits relations (denoted by the '$\sim$' symbol) to be detected between two different 'folds', i and j (i.e. $d_i \sim d_j$), at the sub-domain level, without requiring that the relation also be preserved with the stringency of matched topologies at the higher 'level' of the full domain. That is, $d_i \sim d_j \nRightarrow \mathcal{D}_i \sim \mathcal{D}_j$, even though $d_i \subset \mathcal{D}_i$ and $d_j \subset \mathcal{D}_j$ (in contrast to how patterns of protein structural similarity are traditionally conceived, at the domain level). Here, we can view the characteristic stringency or 'threshold' level of the Urfold, 'd', as being near that of architecture, while $\mathcal{D}$ reflects both architecture *and* topology (corresponding to the classical usage of the term 'fold').

Latent Spaces Capture Gross Structural Properties Across Many Superfamilies, and Reveal a Highly Continuous Nature of Fold Space

The latent space of each superfamily-level DeepUrfold model offers a new, nuanced view of that superfamily, and examining the patterns of similarities among such models may offer a uniquely informative view of fold space. Each SF-specific model captures the different 3D geometries and physicochemical properties that characterize that individual SF as a single 'compressed' data point or embedding; in this way, the latent

[4] In the sense that DeepUrfold's likelihood-based scores can be viewed as measures of the goodness-of-fit of protein domain structures to VAE models that are learnt, against the variational objective, at the SF level.

space representation (or 'distillation') is more comprehensible than is a full 3D domain structure (or superimpositions thereof). In a sense, the DeepUrfold approach—and its inherent latent space representational model of protein SFs, with featurized proteins—can reconcile the dichotomy of a continuous versus discrete fold space because the Urfold model (i) begins with no assumptions about the nature of fold space (i.e., patterns of protein interrelationships), and (ii) does not restrictively enforce full topological ordering as a requirement for a relation to be detected (even a rather weak one) between two otherwise seemingly unrelated domains (e.g., $d_i^{\mathsf{SH3}} \sim d_j^{\mathsf{OB}}$ is not forbidden, using the terminology introduced above). We posit that DeepUrfold can detect these weak similarities (i.e., exhibit high sensitivity) because it operates on protein domains that are featurized beyond purely 3D spatial coordinates; our rationale here is that molecular evolution acts on proteins holistically, not on merely their 3D geometries.

As a first view of fold space through the lens of the Urfold, we used DeepUrfold to represent/compute and analyze the latent spaces of representative domains for highly populated SFs, including mapping the latent space embeddings into two dimensions (Fig 3). Proteins that share similar geometries and biophysical properties should have similar embeddings, and would be expected to lie close together in this latent-space representation, regardless of the annotated 'true' SF. Though this initial picture of the protein universe is limited to 20 highly populated CATH SFs (in this work), already we can see that these SF domains appear to be grouped and ordered by secondary structure composition (Fig 3)—a result that is consistent with past analyses which used approaches such as multidimensional scaling to probe the overall layout of fold space (e.g., [20]). Variable degrees of intermixing between SFs can be seen in UMAP projections such as illustrated in Fig. 3; this is a compelling finding, with respect to the Urfold and its relaxed notion of allowing for intermixed superfamilies. In addition to this mixing, the latent space projection is not punctate: rather than consist of clearly demarcated, well-separated 'islands', instead it is fairly 'compact' (in a loose mathematical sense) and well-connected, with only a few disjoint outlier regions. Manual inspection of these outlier domain structures shows that many of them are incomplete sub-domains or, intriguingly, a single portion of a larger domain-swapped region [46]. Together, these findings support a rather continuous view of fold space, at least for these 20 exemplary superfamilies.

While each superfamily model is trained independently, with different domain structures (SH3, OB, etc.), we find that the distributions that the VAE-based SF models each learn—again, as 'good' approximations to the true likelihood, $p_i(x_{ij}|\boldsymbol{\theta_i})$—are similar, in terms of the dominant features of their latent spaces. In other words, the multiple VAE models (across each unique SF) each learn a structurally low-level, 'coarse-grained' similarity that then yields the extensive overlap seen in Fig. 3. When colored by a score that measures secondary structure content, there are clear directions along which dominant features of the latent-space can be seen to follow, as a gradient from 'all-α' domains to 'all-β' domains, separated by 'α/β' domains. These findings are consistent and reassuring with respect to previous studies of protein fold space (e.g., [20]), as well as the geometric intuition that the similarity between two

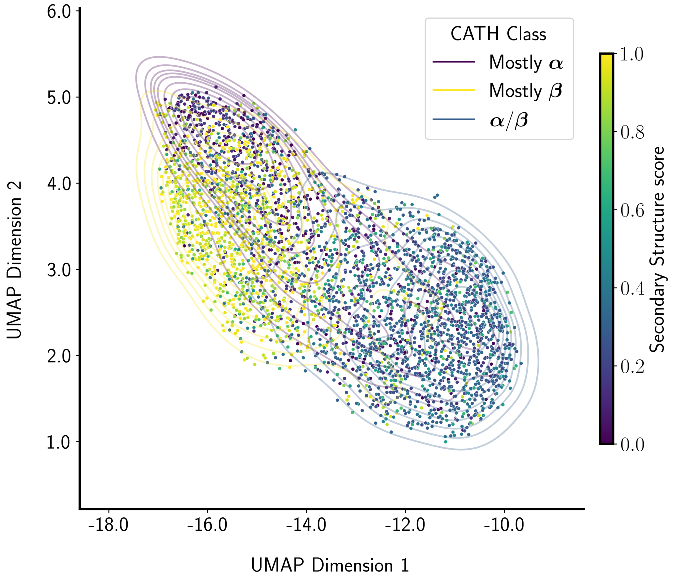

Fig. 3 Dominant variables of DeepUrfold's latent-space models capture gross structural properties and indicate a highly continuous fold space. In a pilot study, we used DeepUrfold to develop 20 distributions/models for 20 CATH homologous superfamilies. Representatives from each SF were subjected to deep models that were trained on domains from the same SF, and then the latent space variables for each structural domain were examined via the uniform manifold approximation and projection (UMAP) method, thereby reducing the 1024 dimensions of the actual model to the two-dimensional projection shown here. In this representation, kernel density estimates (isodensity contour lines) surround domains with the same annotated CATH *Class*. Each domain is colored by a secondary structure score; computed as $(\frac{1}{2}(\#\beta \text{ atoms} - \#\alpha \text{ atoms})/(\#\beta \text{ atoms} + \#\alpha \text{ atoms}) + 0.5)$, this score ranges from zero (for all-α) to unity (for all-β). The protein domains here, as captured in DeepUrfold, can be seen to group together by secondary structure composition; moreover, they are roughly ordered, with the α/β region extensively overlapping the mostly-β region (yellow, predominantly in the vertical direction) and mostly-α region (purple, running predominantly horizontally).

domains would roughly track with their secondary structural content (e.g., two arbitrary all-β proteins are more likely to share geometric similarity than would an all-β and an all-α).

Protein Interrelationships Defy Discrete Clusterings

Our initial finding that fold space is rather continuous, at least under the DeepUrfold model, implies that there are, on average, webs of interconnections (similarities, relationships) between a protein fold $\mathcal{A}$ and its neighbors in fold space ($\mathcal{A}'$, $\mathcal{A}''$, $\mathcal{B}, ...$). Therefore, we posit that an optimally realistic view of the protein universe will not entail hierarchically clustering proteins into mutually exclusive bins, regardless of whether such binning is based upon their folds (giving *fold space*) or any other relatively simple (standalone) geometric feature/criterion. Alternatives to discrete clustering could be such approaches as *fuzzy clustering*, *multi-label classification*, or *mixed-membership community detection* algorithms. DeepUrfold's strategy is to detect communities of similar protein domains, at various levels of stringency, based on the quantifiable similarities of their latent-space representations (versus, e.g., hierarchical clustering based on RMSD or other purely-geometric measures). Again, this is possible because we are armed with a battery of ELBO-based scores of the 'fit' of each SF domain representative to each of the top 20 SF VAE models (Fig 1C).

In DeepUrfold, we formulate this labeling/classification/grouping task as a problem in nonparametric Bayesian stochastic block modelling (SBM; [47, 48]). In particular, we fit an edge-weighted [49], degree-corrected, mixed-membership [50, 51], hierarchical [52] SBM to a fully connected bipartite graph that is built from the similarity scores between (i) the VAE-based SF-level models (one side of the bipartite graph) and (ii) representative structural domains from the representative SFs (the other side of the bipartite graph), as schematized in Fig 1C. In our case, we capture the 'fit' between a domain representative and a particular SF (more precisely, that SF's VAE model) by weighting each edge by the quantity $-\log(-(\text{ELBO}))$ (see Fig 1 and Eq 1). The motivation for this approach is that the full, global collection of $-$(ELBO)-weighted protein interrelationships (again, between SFs and domain representatives) most naturally corresponds to a bipartite graph, or network, which can be represented by its adjacency matrix, $\boldsymbol{A}_{d\times sfam}$; this matrix features covariate edge weights $\boldsymbol{x}$ that link vertices from the two 'sides' of the bipartite graph, where $sfam \in$ 20 highly-populated SFs and $d \in$ 3,674 representative domains from the 20 SFs. Following Peixoto [49], we can write the full joint probability of a given bipartite graph/network occurring by chance—with precisely the same vertices connected by the same edges, with the same weights—as the following product over distributions of data and model parameters:

$$\begin{aligned} &P(A,x,\gamma,G,k,e,b) = \\ &\quad P(A|G)P(x|G,\gamma)P(\gamma|e,b)P(G|k,e,b)P(k|e,b)P(e|b)P(b) \end{aligned} \tag{2}$$

where $\boldsymbol{b}$ is the overlapping partition that represents the numbers of blocks (protein communities) and their group memberships (which nodes map to which blocks), $\boldsymbol{e}$ is a matrix of edge counts between the groups (thus allowing for mixed-membership between blocks), $\boldsymbol{k}$ is the labelled degree sequence, and $\boldsymbol{G}$ is a tensor representing the labeling of half-edges (each edge end-point r, s) to account for mixed-membership, satisfying the constraint $A_{ij} = \sum_{rs} G_{ij}^{rs}$. The edge covariate parameters $\boldsymbol{x}$ (e.g., ELBO-based scores) are sampled from a microcanonical distribution, $P(x|G,\gamma)$, where γ imposes a hard constraint such that $\sum_{ij} G_{ij}^{rs} x_{ij} = \gamma_{rs}$ (Sec. VIIC of [47] and personal communication with T. Peixoto). We seek an SBM that best captures $\boldsymbol{A}$, where 'best' is meant as the usual trade-off between model accuracy (to the observed data) and model simplicity (i.e., mitigating overparametrization). An optimal SBM is obtained by considering this as a nonparametric Bayesian inference problem, meaning that (i) model features (the number of groups/blocks, node membership in blocks, patterns of edges between nodes and between groups, etc.), as well as (ii) model parameters and hyperparameters that are sampled over (marginalized out, via integration), are not set *a priori* but rather are determined by the data itself.

We estimate the optimal parameters for a given SBM via Markov chain Monte Carlo (MCMC) methods. Several different models are created for different $\boldsymbol{b}$ and $\boldsymbol{e}$ in order to find the optimal number of blocks with overlapping edges between them, and these are evaluated using a posterior odds-ratio test [50, 51].

Armed with the above SBM methodology, we can now summarize DeepUrfold's overall approach as consisting of the following four stages: (i) Dataset construction, e.g. via the aforementioned discretization of 3D structures and biophysical properties into voxelized representations [44]; (ii) Training of SF-specific models, using our hybrid stacked 3D-CNN/VAE-based deep networks; (iii) In an inference stage, calculation of ELBO-based scores for 'fits' obtained by subjecting SF representative i to the VAE models of all other SFs, j; (iv) To decipher any patterns amongst these scores, utilization of SBM-based analysis of 'community structure' within the complete set of similarity scores for the VAE-based SF-level models (i.e., the full bipartite network, $\mathrm{SF}_i \times \mathrm{model}_j$).

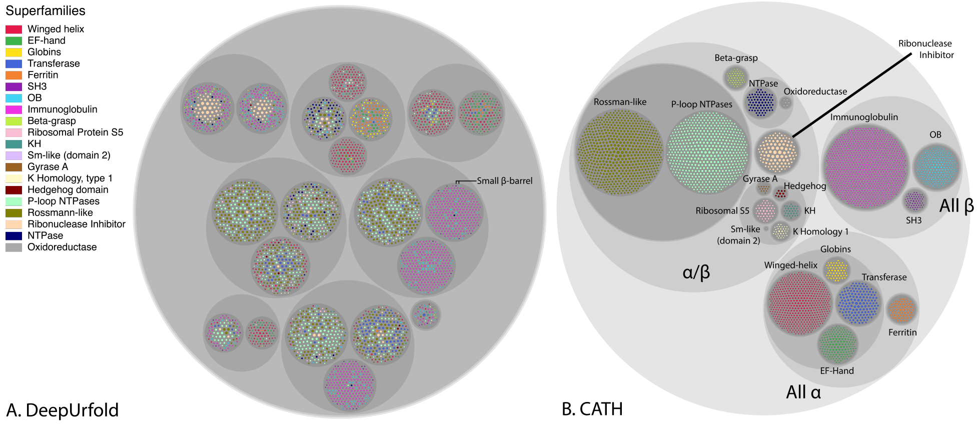

Fig. 4 Protein interrelationships defy discrete clusterings: Stochastic block modeling of an all-vs-all comparison of domain structures and superfamily models. Here, we depict (A) the SBM communities predicted by DeepUrfold as a circle packing diagram, following a similar representational scheme as for (B) the CATH hierarchy. While DeepUrfold avoids hierarchical clustering, we display the groupings in this manner for the sake of visual representation and to facilitate comparison to CATH. Each domain representative is drawn as an innermost circle (corresponding to leaves in a hierarchical tree), colored by the annotated CATH SF and sized by the number of atoms. All of the SF labelled nodes were found to cluster together and were removed from this list (Supp. Fig. 15). Note that many SH3 and OB domains lie within the same lowest-level communities (labeled 'Small β-barrel' in (A)), showing that DeepUrfold can detect the link between these folds, as posited in the Urfold model. Indeed, comparison of the patterns of groupings in (A) to the CATH hierarchy in (B) reveals that DeepUrfold is learning a rather different, non-hierarchical map of protein relationships.

Application of this DeepUrfold methodology to the 20 most highly-populated CATH superfamilies leads us to identify many potential communities of domain structures and SFs (Fig. 4). Subjecting all domain representatives to all 20 SF-specific models, in an exhaustive $all_{\text{SF-models}} \times all_{\text{SF-reps}}$ analysis, reveals the global community structure shown in Fig. 4. We argue that two proteins drawn from vastly different SFs (in the sense of their classification in databases such as CATH or SCOP) can share other, more generalized (e.g., non-contiguous) regions of geometric/structural and biophysical properties, beyond simple permutations of secondary structural elements. And, we believe that the minimally-heuristic manner in which the DeepUrfold model is constructed allows it to capture such 'distant' linkages. In particular, these linkages can be identified and quantitatively described as patterns of similarity in the DeepUrfold model's latent space. Organizing protein domains and superfamilies based on this new similarity metric provides a new view of protein interrelationships—a view that extends beyond simple structural/geometric similarity, towards the realm of integrated *sequence* ↔ *structure* ↔ *function* properties.

We find that domains that have similar −(ELBO) scores against various superfamily models (differing from the SF against which they were trained) are more likely to contain important biophysical properties at particular—and, presumably, functionally important—locations in 3D space; these consensus regions/properties can be thought of as 'defining' the domain.[5] Furthermore, if two domains map into the same SBM community, it is likely that both domains share the same scores when run through each SF model (i.e., an inference calculation), so we hypothesize that that community might contain an urfold that subsumes those two domains (again, agnostic of whatever SFs they are labeled as belonging to in CATH or other databases). We also expect that some domains (those which are particularly 'gregarious'?) may be in multiple communities, which may reflect the phenomenon of a protein being constructed of a multifarious 'urfold' or of several sub-domain elements. Because of the conceptual difficulties and practical complexities of analyzing, visualizing and otherwise representing such high-dimensional data, in the present work we show only the single most likely cluster that each protein domain belongs to, while emphasizing that multi-class membership is a key property of DeepUrfold's approach.

Given the stochastic nature of the SBM calculation, we ran six different replicates. While each replica produced slightly different hierarchies and numbers of clustered communities (ranging from 19-23), the communities at the lowest (coarsest) level remained consistent, and exhibited varying degrees of intermixing. Notably, in each of the replicates the SH3 and OB clustered into the same communities, and likewise the Rossman-like and P-loop NTPases did too, instead of exclusively occupying their own individual clusters; this finding is consistent with the Urfold view of these SFs, as predicted based on manual/visual analysis [12]. In Fig. 4, we chose to display the replica with 20 SFs and highest overlap score compared to CATH in order to enable easy comparison to and reference to CATH. Most notably, each community contains domains from different superfamilies (Fig. 4A), consistent with the Urfold model of protein structure. In the particular subset of proteins treated here, the domains from 'mainly α' and 'α/β' are preferentially associated, while domains from 'mainly β' and 'α/β' group together (Fig. 4B); members of the SH3 and OB superfolds cluster together in the same communities (Fig. 4A), corresponding to the first proposed urfold, the SBB [9].

In addition to coloring each domain (node) by its preexisting CATH superfamily label in circle-packing diagrams, such as that of Fig. 4, we also explored coloring domain nodes by other basic types of properties. These additional properties included: (i) secondary structure type, (ii) average electrostatic potential, (iii) average partial charge, and (iv) enriched gene ontology (GO) terms (Supp Figs S16-21); a navigable, web-based interface for exploring these initial DeepUrfold results is freely available

[5] In some sense, these 'defining regions' may play analogous roles in protein domains as do *tokens* in natural language modeling and generation via large language models such as the generative pre-trained transformers (GPT-n series).

at https://bournelab.org/research/DeepUrfold/. Interestingly, domains with similar average electrostatic potentials (Supp Fig S16) and partial charges (Supp Fig S17) can be found to associate into similar groups in DeepUrfold, whereas the CATH-based circle-packing diagrams, when colored by those same features, have no discernible order or structuring; whether or not this phenomenon stems from any underlying, functionally-relevant 'signal' is a question of interest for further work.

In order to assess how 'well' our DeepUrfold model does, we compare and contrast our clustering results with CATH. However, we emphasize that there is no reliable, objective ground truth for a map of fold space, as there is no universally-accepted, 'correct' description of fold space (and, it can be argued, even 'fold'). Therefore, we cautiously compare our DeepUrfold results to a well-established system, like CATH, with the awareness that these are two conceptually different approaches to representing and describing protein structure relationships and, thus, the protein universe. Indeed, because our model uses a fundamentally different input representation of proteins, intentionally ignoring all topological/connectivity information, we expect that our model will deviate from CATH in terms of clustering-related measures such as *completeness*, *homogeneity*, *silhouette score*, and *partition overlap* [51]. Given all this, approaches that do differ from CATH—versus matching or recapitulating it—can be considered as representing an alternative view of the protein universe. Somewhat counterintuitively, we deem weaker values of our comparison metrics (e.g., less similarity to CATH) as providing stronger support for the Urfold model of protein structure. Simultaneously, we systematically compared how well other, independently-developed sequence– and structure–based models can reconstruct CATH (Fig 5); in so doing, we included a random baseline model as a sort of 'negative control' in gauging the performance of the DeepUrfold framework (Fig. 5 and Supp Info §6.9). Among all these methods, our DeepUrfold approach produces results that are the most divergent from CATH, consistent with DeepUrfold's approach of taking a wholly new view of the protein universe and the domain-level structural similarities that shape it. We also see that many other algorithms, both sequence-based (Fig. 5, left) and structure-based (Fig. 5, right), have difficulty reconstructing CATH (possibly due to extensive manual curation of CATH), but much more closely reproduce it than does our method. We suspect that this largely occurs because of DeepUrfold's intentional, low-level incorporation and integration of more *types* of information than purely 3D structural geometry.

Discussion, Further Outlook

This work offers a new, structure-guided, community-based view of protein relationships. Using a deep learning-enabled framework that we term *DeepUrfold*, our approach aims to (i) explore and assess the Urfold model of protein structure relationships [12], in a rigorous/quantitative manner, and (ii) develop a platform for systematically identifying putative new urfolds. The following are key features of the DeepUrfold framework: (i) It is sensitive to 3D structure and structural similarity between pairs of proteins, but is minimally heuristic (e.g., it does not rely upon pre-set RMSD thresholds or the like) and, crucially, it is topology-agnostic and alignment-free (as it leverages latent space embeddings of featurized structures, versus direct 3D

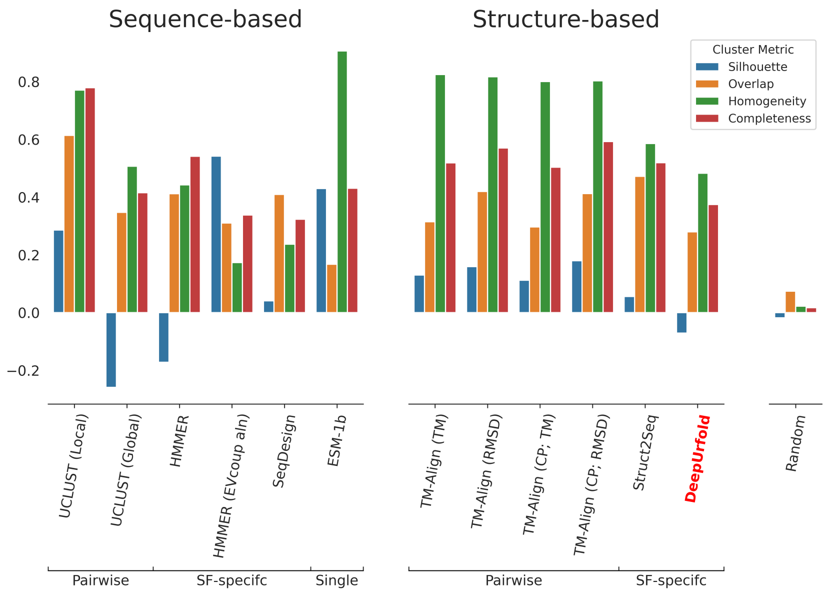

Fig. 5 Comparing DeepUrfold and other methods to CATH. We compare DeepUrfold to other sequence-based (left-half) and structure-based (right-half) protein similarity approaches by using each of them to attempt to reconstruct CATH's organization of protein superfamilies. The scores from each of the algorithms, applied to the same protein dataset as used for DeepUrfold in this work, are used as edge weights to compute an SBM. In so doing, any score types that would increase with decreasing similarity (i.e., correspond to a distance metric) were converted to a similarity metric by negation ($-x$ or $-\log x$). We take the communities at the lowest hierarchical level as clusters and use cluster comparison metrics to understand how well each algorithm/similarity metric can be used to recapitulate CATH. For each of these metrics (*silhouette* value, *overlap*, *homogeneity* and *completeness*), a value of unity is deemed best. DeepUrfold does 'poorly' with these metrics because it does not produce the same clustering patterns—in other words, it is learning something entirely different than are other algorithms, which more closely reproduce CATH. For TM-Align, 'CP' stands for Circular Permutation. We also compared a uniform random grouping for 20 groups as a baseline. For more detailed information, see Supp Info §6.9 and Supp Table S3.

coordinates, for comparison purposes). (ii) Beyond the residue-level geometric information defining a 3D structure (i.e. coordinates), DeepUrfold is an extensible model insofar as it can incorporate *any* types of properties of interest, so long as such data can be encoded as part of the 'featurization' in a deep model—e.g. biophysical and physicochemical characteristics (electrostatic charge, solvent exposure, etc.), site-by-site phylogenetic conservation, and so on. (iii) The DeepUrfold method provides a quantitative metric, in the form of the deep neural network's loss function (at the inference stage), that is amenable to approaches that are more generalized than brute-force hierarchical clustering; for instance, this work shows that we can use loss function scores in stochastic block modeling to construct mixed-membership communities of proteins. In the above ways, DeepUrfold can be viewed as an integrative approach that, while motivated by structural (dis)similarities across fold space, is also cognizant of $\textit{sequence} \leftrightsquigarrow \textit{structure} \leftrightsquigarrow \textit{function}$ interrelationships—even those which are quite weak.

This is intentional: molecular evolution acts on the sequence/structure/function triad as its base 'entity', not on the purely geometric aspects of 3D structure alone. We suspect that any purely geometric/structure-based approach will be limited in its ability to accurately represent fold space (as also described in Supp Info §5.6).

Using the DeepUrfold methodology, we demonstrate (i) the general utility of a new type of similarity metric for representing and comparing protein domain structures, based on deep generative models and latent spaces, and (ii) that a mixed-membership community detection algorithm can identify what we previously found, via manual/visual analysis [12], to be putative urfolds. Finally, we emphasize that because DeepUrfold is agnostic of precise protein topology (i.e., order of connectivity of SSEs in 3D-space), it can readily detect levels of similarity 'above' the fold level (above CATH's 'T' level, below its 'A' level), including the potential of non-contiguous fragments. We believe that such spatially-compact groups of frequently recurring subdomain fragments, sharing similar architectures (independent of topology) within a given group—which, again, we term an 'urfold'—could correspond to primitive 'design elements' in the early evolution of protein domains [22]. We note that Kolodny [53] has made similar points.

Overall, the DeepUrfold framework provides a sensitive approach to detect and thus explore distant protein interrelationships, which we suspect correspond to weak phylogenetic signals (perhaps as echoes of remote/deep homology). Also notable, the embeddings produced by our VAE models and ELBO-based similarity scores provide new methods to visualize and interpret protein interrelationships on the scale of a full fold space. From these models, it is clear that there is a fair degree of continuity between proteins in fold space, and intermixing between what has previously been labeled as separate superfamilies; a corollary of this finding is that discretely clustering proteins, or their embeddings, is ill-advised [54] from this perspective of a densely-populated, smoother-than-expected fold space. An open question is the degree to which the extent of overlap between individual proteins (or groups of domains, as an urfold) in this fold space is reflective of underlying evolutionary processes, e.g. akin to Edwards & Deane's finding [21] that "evolutionary information is encoded along these structural bridges [in fold space]".

While the present work focused exclusively on developing DeepUrfold with CATH as a backdrop, it also would be intriguing to assess other classification schemes as contexts for DeepUrfold-based VAE models—specifically, SCOP, SCOP2 and ECOD. SCOP2 is particularly interesting because it aims to represent sub-domain-level similarities and evolutionarily-distant functional relationships by relaxing the strict constraints of hierarchical trees in favor of a graph-based approach to relationships [33]. A comparative analysis of DeepUrfold groupings (e.g., from the SBM) and SCOP2 groupings, in order to gauge any clear and easily identifiable points of concordance between these approaches, would be of great interest.

Another informative next step would be to use DeepUrfold to identify structural fragments that contain similar patterns of geometry and biophysical properties between proteins from quite different superfamilies. Notably, these fragments may be continuous or discontinuous, and pursuing this goal might help unify the 'primordial peptides' [2] and 'themes' [27] concepts with the Urfold hypothesis, allowing connections between unexplored (or at least under-explored) regions of fold space. Also, we suspect that 'Explainable AI' techniques, such as layer-wise relevance propagation (LRP; [55, 56]), can be used to elucidate which atoms/residues, along with their 3D locations and biophysical properties, are deemed most important in defining the various classification groups that are identified by DeepUrfold (i.e., the structural and physicochemical determinants of why a given protein falls into urfold $\mathcal{A}$ versus urfold $\mathcal{B}$). This goal can be pursued within the DeepUrfold framework because we discretize full domain structures into voxels as part of the 3D-CNN data encoding scheme (Supp Info §2): thus, we can probe the neural network (i.e., trained model) to learn about specific voxels, or groups of specific voxels (e.g., amino acid residues), that contribute as sub-domain structural elements. Doing so would, in turn, be useful in finding common sub-domain segments from different superfamilies. We hypothesize that the most 'relevant' (in the sense of LRP) voxels would highlight important sub-structures; most promisingly, that we know the position, biochemical and biophysical properties, and so on about the residues would greatly illuminate the *physical* basis for the deep learning-based classification. In addition, this would enable us to explore in more detail the mechanistic/structural basis for the mixed-membership features of the SBM-based protein communities. Beyond helping to detect and define new urfolds, for use in areas like protein engineering or drug design, such communities of weakly-related proteins may offer a powerful new lens on remote protein homology.

Online Methods

The following subsections describe the computational methodology that underlies the DeepUrfold framework.

Datasets

Using 'Prop3D', a computational toolkit that we have developed for handling protein properties in machine learning and structural bioinformatics pipelines [44], we created a 'Prop3D-20sf' dataset. This dataset employs 20 highly-populated, diverse CATH superfamilies of interest (Fig 1C); these superfamilies are enumerated in Supp Table S1, which includes annotated rationales for many of the SFs (in the table and its accompanying text). Domain structures from each of the 20 SFs are 'cleaned' by adding missing residues with MODELLER [57], missing atoms with SCWRL4 [58], and protonating and energy minimizing (simple debump) with PDB2PQR [59]. Next, we compute a host of derived properties for each domain in CATH [44], including (i) purely geometric/structural quantities, e.g. secondary structure labels [60] and solvent accessibility, (ii) physicochemical properties, e.g. hydrophobicity, partial charges, electrostatic potentials, (iii) basic chemical descriptors (atom and residue types), and (iv) phylogenetic conservation. As detailed in [44], these computations rely heavily on the Toil workflow engine [61], and data were stored using the Hierarchical Data Format (version 5) in the Highly Scalable Data Service (HSDS). The domains from each SF were split such that all members of an S35 35% sequence identity cluster (pre-calculated by CATH) were on the same side of the split; as described in [44], we constructed data splits so as to mitigate evolutionary 'data leakage'. We partition the protein data roughly as 80% training, 10% validation, and 10% test (https://doi.org/10.5281/zenodo.6873024; further technical details regarding enactment of the computational workflows can be found in [44]).

In our Prop3D dataset, each atom is attributed with the following seven groups of features, which are one-hot (Boolean) encoded: (i) Atom Type (C, CA, N, O, OH, Unknown); (ii) Residue Type (ALA, CYS, ASP, GLU, PHE, GLY, HIS, ILE, LYS, LEU, MET, ASN, PRO, GLN, ARG, SER, THR, VAL, TRP, TYR, Unknown); (iii) Secondary Structure (Helix, Sheet, Loop/Unknown); (iv) Hydrophobic (or not); (v) Electronegative (or not); (vi) Positively-charged (or not); and (vii) Solvent-exposed (or not). For all of the DeepUrfold final production models reported here, the "residue type" feature was omitted because it was found to be uninformative, at least for this type of representation (see Supp Info §3 and Supp Figs S3-4); interestingly, this finding about the dispensability of a residue-type feature was presaged in early work on this project (e.g., the receiver operating characteristic (ROC) curves in Fig 2 of ref [62]).

Protein 3D Structure Representation

We represent protein domains in DeepUrfold's 3D-CNN by discretization into 3D volumetric pixels, or voxels, as described in Supp Info §2. Briefly, our method centers a protein domain in a 256^3 Å^3 cubic volume to allow for large domains, and each atom is mapped to a 1Å^3 voxel using a kD-tree data structure, with a query ball radius set to the van der Waals radius of the atom from a lookup table. If two atoms occupy the same given voxel—a possibility, as the solid diagonal of such a cube is $\sqrt{3} \approx 1.732$Å—then the maximum (element-wise) between their feature vectors is used for that voxel (justifiable because they are all binary-valued). Because a significant fraction of voxels in our representation domain do not contain any atoms, protein domain structures can be encoded in this way via a sparse representation; doing so, via an implementation using MinkowskiEngine [63], substantially reduces the computational costs of our deep learning workflow.

Because there is no unique or 'correct' canonical orientation of a protein structure in $\mathbb{R}^3$, we applied random rotations to each protein domain structure as part of the model training routine; these rotations were in the form of orthogonal transformation matrices randomly drawn from the Haar distribution, which is the uniform distribution on the 3D rotation group, i.e., SO(3) [64].

Stacked 3D-CNN/VAE Model Design and Training

A sparse 3D-CNN variational autoencoder was adapted from MinkowskiEngine [63, 65]. In DeepUrfold's Encoder, there are seven blocks consisting of Convolution (n->2n), BatchNorm, Exponential Linear Unit (ELU) activation functions, Convolution (2n->2n), BatchNorm, and ELU, where n=[16, 32, 64, 128, 256, 512, 1024], or a doubling at each block. Finally, the tensors are pooled using a Global Pooling routine, and the model outputs both a normal distribution's mean and log variance. Next, the learned distribution is sampled from[6] and used as input to the Decoder. The decoder also consists of seven blocks, where each block consists of ConvolutionTranspose(2n->n), BatchNorm, ELU, Convolution(n->n), BatchNorm, and ELU. Finally, one more convolution is used to output a reconstructed domain structure in a 256^3 Å^3 volume. A detailed layout of DeepUrfold's model architecture can be found in Supp Info §8.

In VAE training calculations, a well-established 'reparameterization trick' enable gradients to be computed for backpropagation steps despite the VAE's latent space variables being sampled stochastically. This is achieved by making only the mean (μ) and variance (σ) differentiable, with a random variable that is normally distributed ($\mathcal{N}(0, \mathbf{I})$). That is, the latent variable posterior $\mathbf{z}$ is given by $\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \bigodot \mathcal{N}(0, \mathbf{I})$, where $\bigodot$ denotes the Hadamard (element-wise) matrix product and $\mathcal{N}$ is an 'auxiliary noise' term [66].

We optimize against the negative Evidence Lower BOund (–(ELBO)) described in Equation 1, which combines into a single quantity (i) the mean squared error (MSE) of the reconstructed domain and (ii) the difference between the learned distribution and the true distribution of the SF (i.e., the KL divergence, or relative entropy between the true/underlying distribution of the data given a model, p, and our learned/inferred posterior distribution of latent parameters given the data, q [66]).

We used stochastic gradient descent (SGD) as the optimization algorithm for parameter updates during NN model training, with a momentum of 0.9 and 0.0001 weight decay. We began with a learning rate of 0.2 and decreased its value by 0.9 every epoch using an exponential learning rate scheduler. Our final network has ≈110M parameters in total and all the networks were trained for 30 epochs, using a batch size of 255 (Supp Fig S2 provides an illustrative example of model training with Igs). We utilized the open-source frameworks PyTorch [67] and PytorchLightning [68] to simplify training and inference, and to make the models more reproducible.

To optimize/tune hyperparameters for DeepUrfold's VAE, we used Weights & Biases Sweeps [69] to parameter-scan across the batch size, learning rate, convolution kernel size, transpose convolution kernel size, and convolution stride in the Ig model, while minimizing the –(ELBO). We used a Bayesian Optimization search strategy and 'hyperband' method with three iterations for early termination. We found no significant changes to parameters, and therefore used the following default values: convolution kernel size of 3, transpose convolution kernel size of 2, and convolution stride of 2.

Due to a large-scale class imbalance between the number of domains in each superfamily (e.g., over-representation of Igs), we follow the "one-class classifier" approach, creating one VAE for each superfamily. As part of our 'control experiments', we also

[6] A VAE's modeling/learning of this latent space distribution is what makes it a form of *generative* modeling: were one so inclined, the learnt distribution could be used to generate new instances/samples of the type of entity being modeled (a string of text, image data, etc.), in as optimal a manner as possible ('optimal' in terms of the match between statistical distributions of the generated entities relative to the observed data); more concretely, new entities could be created, for example, by interpolating between latent space embeddings. The generative approach contrasts with, e.g., more traditional *discriminative* models, wherein the likelihood of specific labels being associated with specific instances can be assessed and used to classify/discriminate between the different types of instances (versus spawning new ones). A benefit of generative models is that they develop a probabilistic framework that describes the statistics of the observed instance↔label mappings, thus enabling new entities to be created. Such approaches are powerful, e.g., in de novo protein design.

train a joint SH3 and OB model and compare random over- and under-sampling from ImbalancedLearn [42] on joint models of multiple superfamilies (Supp Info §4.2, Supp Fig S8).

All 20 SF models used throughout this work (i.e., in Prop3D-20sf) were trained using between one and four NVIDIA RTX A6000 GPUs on a Lambda Labs Deep Learning workstation.

Evaluation of Model Performance

We calculated the area under the ROC curve (auROC) and the area under the precision-recall curve (auPRC) for the 20 SFs. Representative domains, as defined by CATH, for each superfamily were subjected to their SF-specific VAE models and predicted values were micro-averaged to perform auROC and auPRC calculations. Immunoglobulins were chosen for purposes of display in this work (Supp Info §3, Supp Figs S2-6), and the results for all SFs can be found in the extended Supp Info. All SFs resulted in roughly similar metrics for each of the seven different groups of encoded features (Supp Figs S3-4).

Assessment of the Urfold Model's Topological Sensitivity by Systematically Scrambling Loops

To gauge the sensitivity of our DeepUrfold model to loop orderings (i.e., topology), we subjected artificial protein structures, with systematically permuted secondary structural elements, to superfamily-specific VAEs. To do this, we generated a series of fictitious folds by implementing a multi-loop permutation algorithm [43], allowing us to systematically 're-wire' the SSEs found in representative SH3 and OB domains in order to exhaustively sample all possible topological orderings (numbering 96, in the case of the SBB's 4-stranded β-sheet). We stitched together the SSEs in various orders and relaxed the conformations/energetics of each new 3D structure using the MODELLER suite [57].

Next, each novel permuted structure is subjected to a VAE model trained on all other domains from the SH3 homologous superfamily. Fit to the model is approximated by the log-likelihood score of the permuted and natural (wild-type) protein represented −(ELBO) scores, which can be viewed as a similarity metric (goodness-of-fit of a given structure to the VAE model). We also calculated a 'background' distribution of each model by performing an all-vs-all TM-align calculation for all domains in our representative CATH domain set; in this step, we recorded any domains that have a TM-score ≤ 0.3, as that threshold quantity is thought to correspond to domains that have random 3D structural similarity (see also the description in Supp Info §4).

Exploration of Latent-space Organization

We subjected all representative domains (numbering 3,674) from each individual SF to an inference pass through each of the 20 SF-specific DeepUrfold models, and visualized the 20 different latent space embeddings for all representatives from each separate model. These results are further detailed in Supp Info §5: in particular, Supp Info §5.4

describes the individual, SF-level feature embeddings that we analyzed as 20 independent subspace projections (Supp Fig S12 shows each of these). More concretely, a 'latent space' for a given domain from one SF-specific VAE corresponds to a 1,024-dimensional vector describing the representative domain in its most 'compressed' or 'distilled' form in the feature space learned by the VAE model, accounting for the position of each atom, their biophysical properties (represented by the mean of the learned distribution), and any other features that were included in the model (e.g., phylogenetic properties; see above). We then pooled the latent spaces for every domain from each superfamily-specific VAE into a single dataset by concatenating on the feature or column dimension, e.g. the shape of the dataset from a single superfamily model is (3674, 1024) and the combined dataset becomes (3674, 20480) after concatenation.

The DeepUrfold-learned embeddings from two different, independently-trained SF-specific VAE models may not be directly comparable, as they can in general occupy different regions of the learned latent (hyper)space. This, in turn, makes it problematic to simply concatenate such embeddings in the feature dimension. An alternative approach could be to 'shift' the embedding vectors to a common region of latent space, via a method known as Optimal Transport (OT) for domain adaptation. As shown in Supp Fig S13 and detailed in its accompanying caption, we applied the OT algorithm (using Sinkhorn-based transport with group LASSO L1L2 regularization) and then concatenated on the feature dimension; reassuringly, this process achieved similar results as our more naive concatenation approach, inasmuch as SFs exhibited a clear dispersal in terms of SSE content (i.e., the non–OT-based approach [Fig 3] and OT-based approach [Supp Fig S13] are roughly similar).

Finally, we reduced the number of latent space dimensions to two (giving a (3674, 2)-sized matrix across all domains) in order to aid visualization of the learned embeddings. We achieved this via three dimensionality-reduction approaches, including the uniform manifold approximation and projection (UMAP) method. As a subspace projection method, the UMAP algorithm is more powerful than the principal component analysis (PCA) method, the latter of which assumes linearity in the data (we also applied PCA to the DeepUrfold embeddings [Supp Fig S11]). Also, UMAP more robustly captures long-range structure/correlations in a dataset than does the common t-distributed stochastic neighbor embedding (t-SNE) approach, which we also applied to the DeepUrfold embeddings (Supp Fig S10). Given our naivety about DeepUrfold's latent spaces, we utilized UMAP as a de facto projection approach because it provides both (i) a well-formed metric notion of local distances (e.g., within-clusters) and (ii) better preserval (versus t-SNE) of the topological structure/relationships amongst more distant points in a dataset, e.g., more global-scale, between-cluster orderings.

Mixed-membership Community Detection via SBMs

We performed all-vs-all comparisons of domains and superfamilies by subjecting representative protein domain structures from each of the 20 chosen SFs through each SF-specific one-class VAE model. The −(ELBO) loss score for each (i, j) pair $(\text{domain}_i^{rep}, \text{SF}_j^{model})$ can be used to quantitatively evaluate pairwise 'distances' between SFs by treating the complete set of distances as a fully connected bipartite graph between domains i (one side of the graph) and SF models j (other side of the graph), defined by adjacency matrix $\boldsymbol{A}_{ij}$, with edges weighted by the $-\log(-(\text{ELBO}))$ scores from the covariate matrix, $\boldsymbol{x}$. Stochastic Block Models (SBM; [48]) offer a generative, nonparametric Bayesian inference-based approach for community detection in random graphs [49]. Therefore, we used SBM algorithms to partition the DeepUrfold-derived bipartite graph into communities of domains that have similar distributions of edge covariates between them. Using the SBM likelihood equation (Equation 2), inference is done via the posterior:

$$P(b, G|A, x) = \frac{P(A, x, \gamma, G, k, e, b)}{P(A, x)} \tag{3}$$

where $\boldsymbol{b}$ is the overlapping partition, $\boldsymbol{e}$ is the matrix of edge counts between groups, $\boldsymbol{k}$ is the labelled degree sequence, and $\boldsymbol{G}$ is a tensor representing half-edges (each edge end-point r, s) to account for mixed-membership, satisfying $A_{ij} = \sum_{rs} G_{ij}^{rs}$. Edge covariates $\boldsymbol{x}$ are sampled from a microcanonical distribution, $P(x|G, \gamma)$, where γ adds a hard constraint such that $\sum_{ij} G_{ij}^{rs} x_{ij} = \gamma_{rs}$ (personal communication with T. Peixoto and Sec. VIIC in [47]).

Using the same SBM approach as we did for post-processing the DeepUrfold-derived data (i.e., ELBO-quantified fits between domain representatives and SF-specific VAE models), we also compared our results to community analyses of data that we performed by using state-of-the-art sequence– and structure–based methods for comparing proteins (e.g., HMMER, ESM, SeqDesign, etc. listed in Fig 5 and Supp Table S3). All SBMs were created using fully-connected $n \times m$ bipartite graphs, linking n CATH S35 domains to m SF models. In our current work, we used 3,674 representative CATH domains from 20 superfamilies, yielding a 3,674 × 20-element similarity matrix for each of the various methods (UCLUST, HMMER, SeqDesign, etc.) that we sought to compare. Each SBM was degree-corrected, overlapping, and nested and fit to a real normal distribution of edge covariates. For those methods that give decreasing scores with increasing similarity (i.e., closer to zero is greater similarity), we $-\log$–transformed each score, whereas values from methods with a non-inverse relationship between the score metric and inferred similarity (i.e., higher values mean greater similarity) were unaltered.

While only 'superfamily-specific' methodologies/models would be directly comparable to the task performed by DeepUrfold (e.g., where $n \times m$ matrices are the original output created by subjecting n CATH representative domains without labels to m SF-specific models), for purposes of comparison we also included 'pairwise' and 'single model' methods (Fig. 5). This was accomplished in the following way: For pairwise approaches, an all-vs-all $n \times n$ similarity matrix was created and then converted to $n \times m$ by taking the median distance of a single CATH domain to every other domain in a given SF. What we are calling 'single model' approaches here are those wherein a single model is trained on all known proteins and outputs a single embedding score for each domain, creating an $n \times 1$ vector. To convert that data form into an $n \times m$ matrix, we took the median distance of a single CATH domain embedding to every other domain embedding from a given SF.

Evaluating SBM Communities, and Comparing to CATH

Because we have no ground truth for the new Urfold view of protein structure similarities (and the resultant protein universe), we applied cluster comparison metrics to evaluate each SBM community both in a self-contained manner and as referred against the original CATH clusterings. The specific measures we considered include the following *silhouette score*, *partition overlap*, *homogeneity*, and *completeness*, for each of the various protein comparison approaches listed in Fig. 5:

- **Silhouette Score:** Provides a measure of how similar an object is to its own cluster (*cohesion*) compared to next-closest cluster (*separation*), with values ranging from -1 (poor grouping) to 1 (ideal).
- **Overlap:** Describes the maximum overlap between partitions, by solving an instance of the maximum weighted bipartite matching problem [51].
- **Homogeneity:** The optimal value (1) occurs when each cluster contains only members of a single class; this metric ranges from $[0, 1]$.
- **Completeness:** Ranging from $[0, 1]$, the optimal value (1) occurs when all members of a given class are (presumably correctly) assigned to the same cluster.

All of our comparisons start by using the sequence and structure representatives from CATH's S35 cluster for each of the 20 superfamilies of interest. The code USEARCH [70] was run twice with parameters `-allpairs_local` and `-allpairs_global`; both runs included the `-acceptall` parameter. HMMER [71] models were built using (1) MUSCLE [72] alignments from CATH's S35 cluster; and (2) a deep MSA created from EVcouplings [73] using jackhmmer [71] and UniRef90 of the first S35 representative for each superfamily. Each HMMER model was used to search all representatives, reporting all sequences with bitscores $\geq -10^{12}$. SeqDesign [74] was run using the same MSAs from EVcouplings. Finally, we also compared our DeepUrfold results against the ESM pre-trained protein language model [75].

For other structure-based comparisons, we ran TM-Align [76] on all representative domains, with and without allowing for circular permutations, and saving the RMSD and TM-score values. Struct2Seq [77] was executed with default parameters after converting domain structure representatives into dictionaries in order to match the required form of input.

Finally, as a baseline, we also compare random groupings to CATH. First, we create a uniform random grouping with 20 groups using numpy's `random.choice` function. Next, we tried using the same SBM clustering above using random weights with numpy's `random.rand` function. The random SBM converged into a solution with only two groups: one for all domains and another for all VAE models (Supp Fig S23).

Acknowledgements

We thank Luis Felipe Murillo and John Readey for support with HSDS, as well as Jaime Iranzo and Tiago Peixoto for discussions regarding SBMs; we also thank Menuka Jaiswal, Saad Saleem, and Yonghyeon Kweon for early efforts on this project. Portions of this work were supported by the University of Virginia and NSF Career award

MCB-1350957. EJD was supported by a University of Virginia Presidential Fellowship in Data Science.

Supplementary information. Supplemental information attached as PDF

Declarations

- **Funding.** Portions of this work were supported by the University of Virginia and NSF Career award MCB-1350957. EJD was supported by a University of Virginia Presidential Fellowship in Data Science.
- **Conflict of interest/Competing interests** None declared.
- **Ethics approval**. Not Applicable.
- **Consent to participate**. Not Applicable.
- **Consent for publication**. All author's consent to publication.
- **Availability of data and materials**. The Prop3D framework to create, share and load datasets *and* its associated Prop3D-20sf pre-built dataset are available at https://prop3d.readthedocs.io/. Prop3D contains instructions for connecting to the public Prop3D-20sf HSDS endpoint (http://prop3d-hsds.pods.uvarc.io/about) and parsing the Prop3D-20sf raw hdf5 file (https://doi.org/10.5281/zenodo.6873024). The extended supplemental material, including the 20 pre-trained SF models and raw output from the stochastic block modelling of DeepUrfold and other tools used to compare against can be found at https://doi.org/10.5281/zenodo.6916524. We also provide an accompanying website to explore the SBM communities and the CATH hierarchy at https://bournelab.org/research/DeepUrfold/.
- **Code availability.** All code to build datasets and train models can be found at http://github.com/bouralab/Prop3D and http://github.com/bouralab/DeepUrfold, respectively.
- **Author contributions.** EJD designed and implemented DeepUrfold. CM and SV developed the initial Urfold model. EJD and CM led the manuscript preparation, and CM and PEB advised the project.

References

[1] Kolodny, R., Pereyaslavets, L., Samson, A. O. & Levitt, M. On the universe of protein folds. *Annual Review of Biophysics* **42**, 559–582 (2013).

[2] Alva, V., Söding, J. & Lupas, A. N. A vocabulary of ancient peptides at the origin of folded proteins. *eLife* **4**, e09410 (2015). URL http://dx.doi.org/10.7554/{eLife}.09410.

[3] Kolodny, R., Nepomnyachiy, S., Tawfik, D. S. & Ben-Tal, N. Bridging themes: short protein segments found in different architectures. *Molecular Biology and Evolution* **38**, 2191–2208 (2021). URL http://dx.doi.org/10.1093/molbev/msab017.

[4] Bromberg, Y. *et al.* Quantifying structural relationships of metal-binding sites suggests origins of biological electron transfer. *Sci. Adv.* **8** (2022). URL https://www.science.org/doi/10.1126/sciadv.abj3984.

[5] Alvarez-Carreño, C., Gupta, R. J., Petrov, A. S. & Williams, L. D. Creative destruction: New protein folds from old. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2207897119 (2022).

[6] Youkharibache, P. Protodomains: Symmetry-related supersecondary structures in proteins and self-complementarity. *Methods in Molecular Biology* **1958**, 187–219 (2019). URL http://dx.doi.org/10.1007/978-1-4939-9161-7_10.

[7] Elofsson, A. Progress at protein structure prediction, as seen in casp15. *Current Opinion in Structural Biology* **80**, 102594 (2023). URL https://www.sciencedirect.com/science/article/pii/S0959440X23000684.

[8] Wodak, S. J., Vajda, S., Lensink, M. F., Kozakov, D. & Bates, P. A. Critical assessment of methods for predicting the 3d structure of proteins and protein complexes. *Annual Review of Biophysics* **52**, 183–206 (2023). URL https://doi.org/10.1146/annurev-biophys-102622-084607. PMID: 36626764.

[9] Youkharibache, P. *et al.* The small β-barrel domain: A survey-based structural analysis. *Structure* **27**, 6–26 (2019). URL http://dx.doi.org/10.1016/j.str.2018.09.012.

[10] Mura, C., Randolph, P. S., Patterson, J. & Cozen, A. E. Archaeal and eukaryotic homologs of Hfq: A structural and evolutionary perspective on Sm function. *RNA Biology* **10**, 636–651 (2013). URL https://doi.org/10.4161/rna.24538.

[11] Lee, C. & Wilkinson, D. J. A review of stochastic block models and extensions for graph clustering. *Applied Network Science* **4**, 1–50 (2019).

[12] Mura, C., Veretnik, S. & Bourne, P. E. The *Urfold*: Structural similarity just above the superfold level? *Protein Science* **28**, 2119–2126 (2019). URL http://dx.doi.org/10.1002/pro.3742.

[13] Grishin, N. V. Fold change in evolution of protein structures. *Journal of Structural Biology* **134**, 167–185 (2001). URL http://dx.doi.org/10.1006/jsbi.2001.4335.

[14] Kinch, L. N. & Grishin, N. V. Evolution of protein structures and functions. *Current Opinion in Structural Biology* **12**, 400–408 (2002). URL http://dx.doi.org/10.1016/s0959-440x(02)00338-x.

[15] Krishna, S. S. & Grishin, N. V. Structural drift: a possible path to protein fold change. *Bioinformatics* **21**, 1308–1310 (2005). URL http://dx.doi.org/10.1093/bioinformatics/bti227.

[16] Alvarez-Carreño, C., Penev, P. I., Petrov, A. S. & Williams, L. D. Fold evolution before LUCA: Common ancestry of SH3 domains and OB domains. *Molecular Biology and Evolution* **38**, 5134–5143 (2021). URL http://dx.doi.org/10.1093/molbev/msab240.

[17] Sadreyev, R. I., Kim, B.-H. & Grishin, N. V. Discrete-continuous duality of protein structure space. *Current Opinion in Structural Biology* **19**, 321–328 (2009). URL http://dx.doi.org/10.1016/j.sbi.2009.04.009.

[18] Taylor, W. R. Exploring protein fold space. *Biomolecules* **10** (2020). URL http://dx.doi.org/10.3390/biom10020193.

[19] Holm, L. & Sander, C. Mapping the protein universe. *Science* **273**, 595–603 (1996). URL http://dx.doi.org/10.1126/science.273.5275.595.

[20] Hou, J., Jun, S.-R., Zhang, C. & Kim, S.-H. Global mapping of the protein structure space and application in structure-based inference of protein function. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 3651–3656 (2005). URL http://dx.doi.org/10.1073/pnas.0409772102.

[21] Edwards, H. & Deane, C. M. Structural bridges through fold space. *PLoS Computational Biology* **11**, e1004466 (2015). URL http://dx.doi.org/10.1371/journal.pcbi.1004466.

[22] Skolnick, J., Arakaki, A. K., Lee, S. Y. & Brylinski, M. The continuity of protein structure space is an intrinsic property of proteins. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 15690–15695 (2009). URL http://dx.doi.org/10.1073/pnas.0907683106.

[23] Friedberg, I. & Godzik, A. Fragnostic: walking through protein structure space. *Nucleic Acids Research* **33**, W249–51 (2005). URL http://dx.doi.org/10.1093/nar/gki363.

[24] Harrison, A., Pearl, F., Mott, R., Thornton, J. & Orengo, C. Quantifying the similarities within fold space. *Journal of Molecular Biology* **323**, 909–926 (2002). URL http://dx.doi.org/10.1016/s0022-2836(02)00992-0.

[25] Harrison, A. *et al.* Recognizing the fold of a protein structure. *Bioinformatics* **19**, 1748–1759 (2003). URL http://dx.doi.org/10.1093/bioinformatics/btg240.

[26] Goncearenco, A., Shaytan, A. K., Shoemaker, B. A. & Panchenko, A. R. Structural perspectives on the evolutionary expansion of unique protein-protein binding sites. *Biophysical Journal* **109**, 1295–1306 (2015). URL http://dx.doi.org/10.1016/j.bpj.2015.06.056.

[27] Nepomnyachiy, S., Ben-Tal, N. & Kolodny, R. Complex evolutionary footprints revealed in an analysis of reused protein segments of diverse lengths. *Proceedings*

of the National Academy of Sciences of the United States of America **114**, 11703–11708 (2017). URL http://dx.doi.org/10.1073/pnas.1707642114.

[28] Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473 (2019). URL http://dx.doi.org/10.1186/s12859-019-3019-7.

[29] Budowski-Tal, I., Nov, Y. & Kolodny, R. FragBag, an accurate representation of protein structure, retrieves structural neighbors from the entire PDB quickly and accurately. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 3481–3486 (2010). URL http://dx.doi.org/10.1073/pnas.0914097107.

[30] Kolodny, R., Koehl, P., Guibas, L. & Levitt, M. Small libraries of protein fragments model native protein structures accurately. *Journal of Molecular Biology* **323**, 297–307 (2002). URL http://dx.doi.org/10.1016/s0022-2836(02)00942-7.

[31] Durairaj, J., Akdel, M., de Ridder, D. & van Dijk, A. D. J. Geometricus represents protein structures as shape-mers derived from moment invariants. *Bioinformatics* **36**, i718–i725 (2020). URL http://dx.doi.org/10.1093/bioinformatics/btaa839.

[32] Sillitoe, I. *et al.* CATH: expanding the horizons of structure-based functional annotations for genome sequences. *Nucleic Acids Research* **47**, D280–D284 (2019). URL http://dx.doi.org/10.1093/nar/gky1097.

[33] Andreeva, A., Howorth, D., Chothia, C., Kulesha, E. & Murzin, A. G. SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Research* **42**, D310–4 (2014). URL http://dx.doi.org/10.1093/nar/gkt1242.

[34] Fox, N. K., Brenner, S. E. & Chandonia, J.-M. SCOPe: Structural classification of proteins–extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Research* **42**, D304–9 (2014). URL http://dx.doi.org/10.1093/nar/gkt1240.

[35] Cheng, H. *et al.* ECOD: an evolutionary classification of protein domains. *PLoS Computational Biology* **10**, e1003926 (2014). URL http://dx.doi.org/10.1371/journal.pcbi.1003926.

[36] Agrawal, V. & Kishan, R. K. Functional evolution of two subtly different (similar) folds. *BMC Structural Biology* **1**, 1–6 (2001).

[37] Theobald, D. L. & Wuttke, D. S. Divergent evolution within protein superfolds inferred from profile-based phylogenetics. *Journal of Molecular Biology* **354**, 722–737 (2005). URL http://dx.doi.org/10.1016/j.jmb.2005.08.071.

[38] LeCun, Y., Bengio, Y. & Hinton, G. Deep Learning. *Nature* **521**, 436–444 (2015).

[39] Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization (2014). URL https://arxiv.org/abs/1412.6980.

[40] Murphy, K. P. *Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning series)* Illustrated edition edn (The MIT Press, 2012).

[41] Prati, R. C., Batista, G. E. A. P. A. & Monard, M. C. Data mining with imbalanced class distributions: Concepts and methods (2009).

[42] Lemaître, G., Nogueira, F. & Aridas, C. K. Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research* **18**, 1–5 (2017). URL http://jmlr.org/papers/v18/16-365.html.

[43] Dai, L. & Zhou, Y. Characterizing the existing and potential structural space of proteins by large-scale multiple loop permutations. *Journal of Molecular Biology* **408**, 585–595 (2011). URL http://dx.doi.org/10.1016/j.jmb.2011.02.056.

[44] Draizen, E. J., Murillo, L. F. R., Readey, J., Mura, C. & Bourne, P. E. Prop3D: A flexible, Python-based platform for machine learning with protein structural properties and biophysical data. *bioRxiv* (2022). URL https://www.biorxiv.org/content/early/2022/12/30/2022.12.27.522071.

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

[46] Liu, Y. & Eisenberg, D. 3D Domain swapping: As domains continue to swap. *Protein science* **11**, 1285–1299 (2002).

[47] Peixoto, T. P. *Bayesian Stochastic Blockmodeling*, Ch. 11, 289–332 (John Wiley Sons, Ltd, 2019). URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119483298.ch11. https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119483298.ch11.

[48] Peixoto, T. P. Nonparametric Bayesian inference of the microcanonical stochastic block model. *Physical Review E* **95** (2017). URL https://doi.org/10.1103%2Fphysreve.95.012317.

[49] Peixoto, T. P. Nonparametric weighted stochastic block models. *Physical review. E* **97**, 012306 (2018). URL https://link.aps.org/doi/10.1103/{PhysRevE}.97.012306.

[50] Peixoto, T. P. Model selection and hypothesis testing for large-scale network models with overlapping groups. *Physical Review X* **5** (2015). URL https://doi.org/10.1103%2Fphysrevx.5.011033.

[51] Peixoto, T. P. Revealing consensus and dissensus between network partitions. *Physical Review X* **11**, 021003 (2021). URL https://link.aps.org/doi/10.1103/{PhysRevX}.11.021003.

[52] Peixoto, T. P. Hierarchical block structures and high-resolution model selection in large networks. *Physical Review X* **4** (2014). URL https://doi.org/10.1103%2Fphysrevx.4.011047.

[53] Kolodny, R. Searching protein space for ancient sub-domain segments. *Current Opinion in Structural Biology* **68**, 105–112 (2021). URL https://www.sciencedirect.com/science/article/pii/S0959440X20302104.

[54] Bourne, P. E., Draizen, E. J. & Mura, C. The curse of the protein ribbon diagram. *PLOS Biology* **20**, 1–4 (2022). URL https://doi.org/10.1371/journal.pbio.3001901.

[55] Montavon, G., Binder, A., Lapuschkin, S., Samek, W. & Müller, K.-R. in *Layer-wise relevance propagation: An overview* (eds Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K. & Müller, K.-R.) *Explainable AI: interpreting, explaining and visualizing deep learning*, Vol. 11700 of *Lecture notes in computer science* 193–209 (Springer International Publishing, Cham, 2019). URL http://link.springer.com/10.1007/978-3-030-28954-6_10.

[56] Hochuli, J., Helbling, A., Skaist, T., Ragoza, M. & Koes, D. R. Visualizing convolutional neural network protein-ligand scoring. *Journal of molecular graphics & modelling* **84**, 96–108 (2018). URL http://dx.doi.org/10.1016/j.jmgm.2018.06.005.

[57] Eswar, N. *et al.* Comparative protein structure modeling using Modeller. *Current Protocols in Bioinformatics* **Chapter 5**, Unit 5.6 (2006). URL http://dx.doi.org/10.1002/0471250953.bi0506s15.

[58] Krivov, G. G., Shapovalov, M. V. & Dunbrack, R. L. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins* **77**, 778–795 (2009). URL http://dx.doi.org/10.1002/prot.22488.

[59] Dolinsky, T. J. *et al.* PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Research* **35**, W522–5 (2007). URL http://dx.doi.org/10.1093/nar/gkm276.

[60] Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983). URL http://dx.doi.org/10.1002/bip.360221211.

[61] Vivian, J. *et al.* Toil enables reproducible, open source, big biomedical data analyses. *Nature Biotechnology* **35**, 314–316 (2017). URL http://dx.doi.org/10.1038/nbt.3772.

[62] Jaiswal, M. *et al. Deep Learning of Protein Structural Classes: Any Evidence for an 'Urfold'?*, 1–6 (2020).

[63] Choy, C., Gwak, J. & Savarese, S. *4d spatio-temporal convnets: Minkowski convolutional neural networks*, 3075–3084 (2019).

[64] Stewart, G. W. The efficient generation of random orthogonal matrices with an application to condition estimators. *SIAM Journal on Numerical Analysis* **17**, 403–409 (1980). URL https://doi.org/10.1137/0717034.

[65] Gwak, J., Choy, C. B. & Savarese, S. *Generative sparse detection networks for 3d single-shot object detection* (2020).

[66] Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *arXiv* (2013). URL https://arxiv.org/abs/1312.6114.

[67] Paszke, A. *et al.* in *Pytorch: An imperative style, high-performance deep learning library* (eds Wallach, H. *et al.*) *Advances in Neural Information Processing Systems 32* 8024–8035 (Curran Associates, Inc., 2019). URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[68] Falcon, W. *et al.* Pytorchlightning/pytorch-lightning: 0.7.6 release (2020). URL https://doi.org/10.5281/zenodo.3828935.

[69] Biewald, L. Experiment tracking with weights and biases (2020). URL https://www.wandb.com/. Software available from wandb.com.

[70] Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010). URL http://dx.doi.org/10.1093/bioinformatics/btq461.

[71] Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Research* **41**, e121 (2013). URL http://dx.doi.org/10.1093/nar/gkt263.

[72] Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792–1797 (2004). URL http://dx.doi.org/10.1093/nar/gkh340.

[73] Hopf, T. A. *et al.* The EVcouplings Python framework for coevolutionary sequence analysis. *Bioinformatics* **35**, 1582–1584 (2019). URL http://dx.doi.org/10.1093/bioinformatics/bty862.

[74] Shin, J.-E. *et al.* Protein design and variant prediction using autoregressive generative models. *Nature Communications* **12**, 2403 (2021). URL http://dx.doi.o

rg/10.1038/s41467-021-22732-w.

[75] Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences of the United States of America* **118** (2021). URL http://www.pnas.org/lookup/doi/10.1073/pnas.2016239118.

[76] Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research* **33**, 2302–2309 (2005). URL http://dx.doi.org/10.1093/nar/gki524.

[77] Ingraham, J., Garg, V., Barzilay, R. & Jaakkola, T. Generative models for graph-based protein design. *Advances in Neural Information Processing Systems* (2019). URL https://papers.nips.cc/paper/2019/hash/f3a4ff4839c56a5f460c88cce3666a2b-Abstract.html.