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

Eli J. Draizen[1,2*], Stella Veretnik[2], Cameron Mura[1,2*], and Philip E. Bourne[1,2]
[1]Department of Biomedical Engineering,
[2]School of Data Science, University of Virginia, Charlottesville, VA, USA
***Corresponding author:** E-mail: e.draizen@gmail.com and cmura@virginia.edu
Associate Editor: Not Decided

Abstract

Unresolved questions about the discrete/continuous dichotomy of protein fold space permeate structural and evolutionary biology. From protein structure comparison and classification to evolutionary analyses and function prediction, our views of fold space implicitly rest upon many assumptions that impact how we analyze, interpret and come to understand biological systems. Discrete views of fold space categorize similar folds into separate groups; unfortunately, such a 'binning' process inherently fails to capture many remote relationships. While hierarchical databases such as CATH, SCOP, and ECOD represent major steps forward in protein classification, we believe that a scalable, objective and conceptually flexible method that is less reliant upon assumptions and heuristics could enable a more systematic and thorough exploration of fold space and evolutionary-distant relationships. Here, we develop a structure-guided, comparative analysis of proteins, leveraging embeddings derived from deep generative models, which represent a highly-compressed, lower-dimensional space of a given protein and its sequence, structure and biophysical properties. Building upon a recent 'Urfold' model of protein structure, the deep generative approach developed here, termed 'DeepUrfold', 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 are not captured by existing methods.

Key words: protein structure, evolution, deep learning, generative models

Introduction

Much remains unknown about the precise historical trajectory of the protein universe (Kolodny *et al.*, 2013), from (proto-)peptides, to protein domains, to multi-domain proteins (Alva *et al.*, 2015). Presumably, the protein universe—meaning the set of all proteins (known or unknown, ancestral or extent)—did not spontaneously arise with intact, full-sized domains. Rather, smaller, sub-domain–sized protein fragments likely preceded the 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 (Alva *et al.*, 2015; Alvarez-Carreño *et al.*, 2022; Bromberg *et al.*, 2022; Kolodny *et al.*, 2021; Youkharibache, 2019). Our ability to detect common fragments, shared amongst at least two domains, relies on (i) having an accurate similarity metric 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 3D structures.

The advent of deep learning, including the application of such approaches to protein sequence and structure representations, creates a new opportunity to study protein interrelationships in a wholly different manner—namely, via quantitative comparison of 'latent space' representations of a protein as lower-dimensional 'embeddings'; such embeddings can be at arbitrary levels of granularity (e.g., atomic), and can subsume virtually any types of properties (such as amino acid type, physicochemical features such as electronegatitivty, and phylogenetic conservation of the site). The present work explores the idea that viewing protein fold space in terms of latent spaces (what regions are populated, with what densities, etc.) is likely to implicitly harbor deep information about protein interrelationships, over a vast multitude of protein evolutionary timescales.

The traditional approach to examining fold space has been to hierarchically cluster domains via 3D structure comparison, as exemplified in databases such as CATH (Sillitoe *et al.*, 2019), SCOP (Andreeva *et al.*, 2014; Fox *et al.*, 2014), and ECOD (Cheng *et al.*, 2014). Despite being some of the most comprehensive resources available, 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 inability to traverse, e.g. hop from island to island or create 'bridges' between the islands, in the fold spaces presented by these databases implies that some folds have no well-defined or discernible relationships, i.e. we miss the weak or more indeterminate (but nevertheless *bona fide*) signals of remote relationships that link distantly-related folds. In addition to 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 lower (more homologous) levels; what relationships might be detectable if we relax the constraints of strict topological identity?

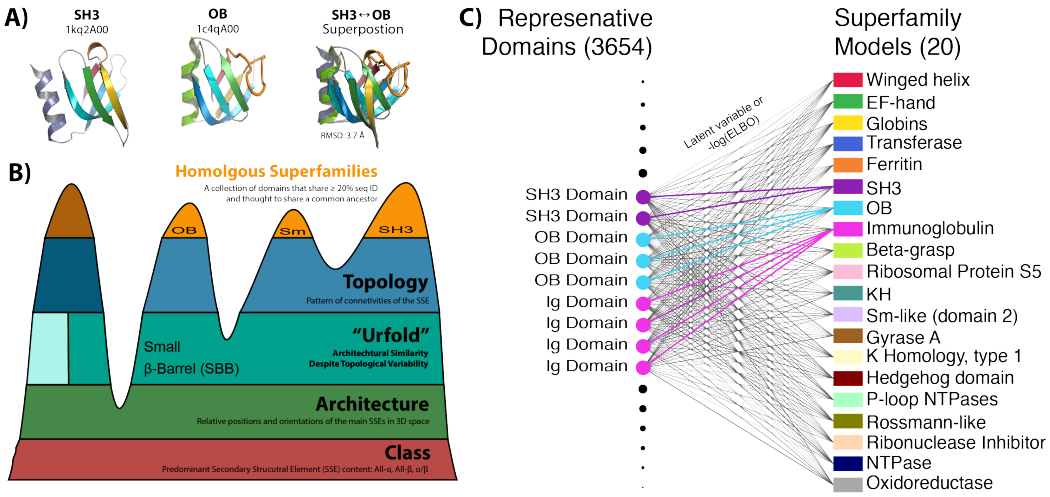

FIG. 1. DeepUrfold method of identifying domains that capture the phenomenon of "architectural similarity despite topological variability". (A) SH3 and OB domains are considered part of the small β-barrel (SBB) 'urfold' because they have the same architecture, yet different topology; they have strikingly similar folds and share many similar functions (e.g. PPI binding on the same edge-strand and nucleic acid binding (Mura *et al.*, 2013; Youkharibache *et al.*, 2019)), yet these similarities are obscured by their having been classified differently. In the case of the SBB urfold, the loops between the β strands have been permuted, resulting in different topologies, observed by their superposition. (B) If the 'Urfold' phenomenon is viewed in terms of CATH, it is hypothesized to be a discrete entity ('level' of structure) that lies between the Architecture and Topology strata. (C) DeepUrfold, which applies deep learning to the Urfold model of protein structure, identifies new potential urfolds by creating 20 superfamily-specific variational autoencoder neural network models and comparing output scores from all representative domains from those superfamilies (3654) to every other superfamily model. For the first metric we compare the latent variables from domain represents through models trained from the same superfamily (colored lines; see Fig. 3) and then we perform an all-vs-all comparison, which is clustered using Stochastic Block Models (see Fig. 4).

The transition of a protein sequence from one fold to another, either naturally (via evolution) or artificially (via design/engineering), likely occurs over multiple intermediate steps, for example by combining and/or permuting short secondary structure segments, *or* mutating single residues (Alvarez-Carreño *et al.*, 2021, 2022; Grishin, 2001; Kinch and Grishin, 2002; Krishna and Grishin, 2005). Each step may correspond to the same or a different fold than its preceding step. The similarity between these transitional states blurs the line of distinct groups—increasing or decreasing a relatively arbitrary and heuristic quantity (namely, the similarity threshold) changes which structures belong to which groups. In this sense, the discrete versus continuous duality of protein fold space can be viewed largely as a matter of semantics or thresholding, versus any 'real' (intrinsic or fundamental) feature of the space itself (Sadreyev *et al.*, 2009).

Pairwise similarity metrics in structure space first indicated remote connections in a continuous fold space via shared fragments (Taylor, 2020). In an early landmark study, (Holm and Sander, 1996) discovered that the protein universe harbors five peptide 'attractors', representing frequently-adopted folding motifs (e.g., the β-meander); this finding rested upon creating an all-by-all similarity matrix from 3D structural alignments. Later, similar pairwise analyses across the protein structure space showed that 'all-α' and 'all-β' proteins are separated by 'α/β' proteins (Hou *et al.*, 2005).

The all-by-all similarity metric of full domains or small fragments can also be viewed as an adjacency matrix of a graph, thereby enabling the creation of a network representation of fold space. Such networks are nearly connected, linking domains in 4-8 steps (Edwards and Deane, 2015; Friedberg and Godzik, 2005; Skolnick *et al.*, 2009).

Graph-based representations of single proteins also pushed the thinking of common short fragments. (Harrison *et al.*, 2002) found maximal common cliques of connected secondary structure elements (SSE) in a graph-based representation of proteins, consisting of SSEs as vertices. In that work, 80% of folds shared common cliques with other folds, and these were quantified as a term they called 'gregariousness'.

Even though short peptide fragments (sub-domain-sized) have been thoroughly studied, relatively few approaches have taken an evolutionary perspective with a continuous fold space. (Goncearenco *et al.*, 2015) identified common loop fragments flanked by SSEs, called Elementary Functional Loops (EFL), that couple in 3D space to perform enzymatic activity. (Youkharibache, 2019) noticed that peptide fragments, called 'protodomains', are often composed (with C2 symmetry) to give a larger, full-sized domain. Most recently, (Bromberg *et al.*, 2022) identified common fragments between metal-binding proteins using *'sahle'*, a new length-dependent structural alignment similarity metric.

The two state-of-the art evolution-based fragment libraries are 'primordial peptides' (Alva *et al.*, 2015) and 'themes' (Nepomnyachiy *et al.*, 2017). Both methods create a set of common short peptide fragments based on HHsearch (Steinegger *et al.*, 2019) profiles for proteins in SCOP and ECOD respectively. The sizes of the libraries created by these two 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 fragments is to describe a protein by a vector of fragments. For example, the FragBag method (Budowski-Tal *et al.*, 2010) describes a protein by the occurrence of fragments in a clustered fragment library (Kolodny *et al.*, 2002). A recent and somewhat unique approach, Geometricus (Durairaj *et al.*, 2020), 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 peptide that duplicated and removed its terminal SSE in a process termed 'creative destruction', resulting in two different folds with different topologies, but similar architecture (Alvarez-Carreño *et al.*, 2021, 2022).

Previously, we identified structure/function-driven connections between several SFs that exhibit architectural similarity despite topological variability, in a new level of structural granularity of discontinuous fragments that we termed the 'Urfold' shown in Fig. 1B (Mura *et al.*, 2019; Youkharibache, 2019). Urfolds were first described in small β-barrel (SBB) domains (Fig. 1A) because of their structure/function similarity in the deeply-varying collection of proteins that adopted either the SH3/Sm or OB superfolds (Youkharibache, 2019). Notably, these 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 (transcription and translation apparatus) throughout all domains of life (Agrawal and Kishan, 2001; Alvarez-Carreño *et al.*, 2021).

Here we present a new method to systematically identify Urfolds using a new alignment-free, biochemically-aware similarity metric of domain structures based on deep generative models and mixed-membership community detection. We leverage similarities in latent-spaces rather than simple/purely-geometric 3D structures directly, and we can encode biophysical and other properties, thereby allowing higher orders of similarity to be detected that may correspond to (dis-)contiguous fragments (Fig. 1C).

Results

Deep generative models can identify similarities between topologically-distinct folds

Conventionally, folds that have similar architectures, but varying topologies, are often thought of as resulting from convergent evolution. However, as in the case with the SH3 and OB superfolds, the structure/function similarities (Youkharibache, 2019), and even sequence/structure/function similarities (Alvarez-Carreño *et al.*, 2021), often prove to be quite striking, suggesting that these domain architectures did not arise independently (Alvarez-Carreño *et al.*, 2021; Youkharibache, 2019). In order to study what may be even quite weak 3D similarities, we model the evolutionary process giving rise to proteins as a 3D structure 'generator'. In so doing, we seek to learn probability distributions $\mathrm{p}(x|\theta)$ that describe the specific geometries and biophysical properties of different folds, 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 parameters describing the variational distribution over the background (i.e., latent) parameters. We posit that folds with similar probabilistic distributions likely have similar geometries/architectures and biophysical properties, regardless of potentially

differing topologies, and that, in turn, may imply a common evolutionary history.

DeepUrfold learns the background distribution parameters $\boldsymbol{\theta_i}$ for 20 superfamily distributions, $\mathrm{p}_i(\mathrm{x}_{ij}|\boldsymbol{\theta_i})$, by constructing variational autoencoders (VAE) for each superfamily i and domain structure j. The original/underlying distribution $\mathrm{p}(\mathrm{x}_{ij}|\boldsymbol{\theta_i})$ is unknown and intractable, so it must be approximated by modeling it as an easier-to-learn distribution, $\mathrm{q}_i(\mathrm{z}_{ij}|\mathbf{x}_i)$. In our case, $\mathrm{q}_i(\mathrm{z}_{ij}|\mathbf{x}_i)$ is taken as sampling from a Gaussian. To ensure $\mathrm{q}_i(\mathrm{z}_{ij}|\mathbf{x}_i)$ can adequately describe $\mathrm{p}(\mathrm{x}_{ij}|\boldsymbol{\theta_i})$, we maximize the Evidence Lower BOund, or ELBO quantity, which is the lower bound of the marginal likelihood of a single structure, $\ln[p_i(x_{ij})]$. The ELBO inequality can be written as:

$$\ln\ \mathrm{p_i}(x_{ij}) \geq E_{q_i(z_{ij}|\mathbf{x}_i)}[\ln\ \mathrm{p}_i(x_{ij}|z_{ij})] - D_{KL}[\mathrm{q}_i(z_{ij}|x_{ij})||\mathrm{p}(z_{ij})] \quad (1)$$

where $\mathrm{p_i}(x_{ij})$ is the log-likelihood, E is the expected value of q in terms of p, and $D_{KL}[p||q]$ is the Kullback-Leibler divergence, or relative entropy, between the two probability distributions q and p. In other words, maximizing the ELBO maximizes the log-likelihood of the model, corresponding to minimizing the entropy between (i) the true underlying distribution $\mathrm{p}_i(\mathrm{x}_{ij}|\boldsymbol{\theta_i})$ and (ii) our learned/inferred posterior distribution of latent parameters given the data, $\mathrm{q}_i(\mathrm{z}_{ij}|\mathbf{x}_i)$. In a similar manner, we train joint models of superfamilies with different topologies, e.g. SH3 and OB, while accounting for the class imbalance (Lemaître *et al.*, 2017; Prati *et al.*, 2009) that stems from there being vastly different numbers of available 3D structural data for different protein superfamilies.

As input to the VAE, we encode the 3D structure of each protein domain by representing it as a 3D volumetric object, akin to the input used in 3D convolutional neural networks (CNNs). In our discretization, atoms are binned into voxels, each of which can be labeled, atom-wise, with arbitrary properties (biophysical, phylogenetic, etc.). This representation 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 the model (and can be used to unambiguously reconstruct the 3D structure).

As an initial assessment of our SH3, OB and joint SH3/OB DeepUrfold models, and to examine the Urfold model more broadly, we explicitly tested the Urfold's underlying concept of "*architectural similarity despite topological variability*". This test was performed by considering artificial protein domains that have identical architectures but specifically introduced loop permutations; this systematic perturbation of a 3D structure's topology was obtained via 'rewiring' of the SSEs (scrambling the loops), while retaining the overall 3D

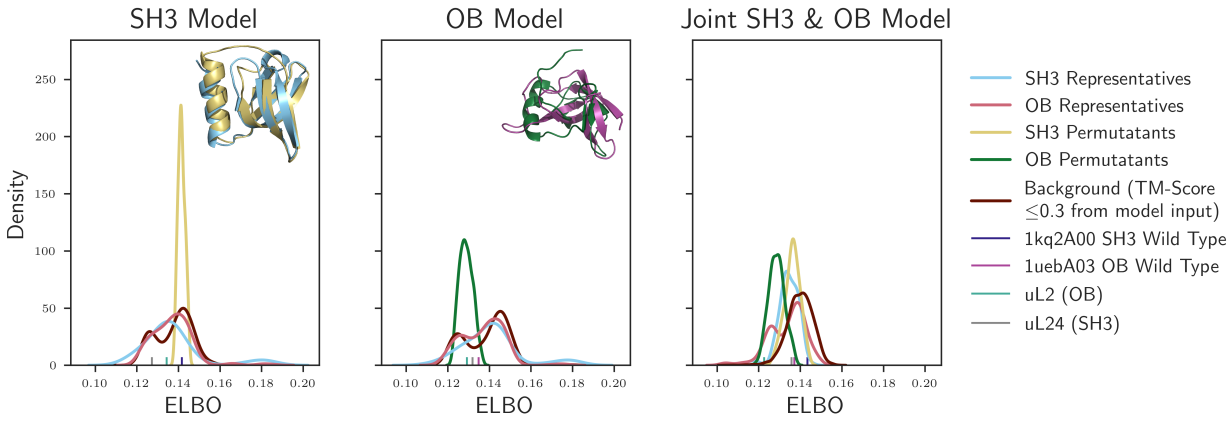

FIG. 2. Likelihood values can be used to quantify similarities among multi-loop permuted structures. To gauge the sensitivity of our DeepUrfold metric to loop orderings (topology) via generation of fictitious folds, we implemented a multi-loop permutation algorithm (Dai and Zhou, 2011) in order to 'scramble' 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 using the MODELLER suite. While 96 unique permutations are theoretically possible for a 4-stranded β-sheet (Youkharibache *et al.*, 2019), 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 model that had been trained on all other domains from either the **(A)** SH3; **(B)** OB; or **(C)** joint SH3/OB models. Fit to the model was approximated by the ELBO score, which can be viewed as a similarity metric or a measure of 'goodness-of-fit'. In reference to a given model, a given permutant query structure having an ELBO score 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 stable. As reference points, we also include the ELBO scores for ancestrally-reconstructed progenitors of the OB (uL2) and SH3 (uL24) superfolds, based on (Alvarez-Carreño *et al.*, 2021).

structure (architecture). Specifically, we (i) created permuted (fictitious) 3D structures for representative SH3 and representative OB domains (Supp. Fig. 7A), and then (ii) subjected these to the SH3, OB, and joint SH3/OB DeepUrfold models. Because small β-barrels (SBBs) typically have six SSEs, including four 'core' β-strands, each β-sheet core of an SBB can theoretically adopt one of at least 96 distinct loop permutations (Youkharibache *et al.*, 2019); 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.

We find that the permuted domain structures have similar ELBO scores as the corresponding wild-type domains (Fig. 2). Those permuted domain structures with ELBO scores *less* than the wild-type domains can be interpreted as being more similar (structurally, biophysically, etc.) to the DeepUrfold variational model, and thus perhaps more thermodynamically stable or structurally robust were they to exist in reality (an interesting possibility as regards protein design and engineering). TM-Scores (Xu and Zhang, 2010) for permuted domain structures against the corresponding wild-type typically lied in the range $\approx 0.3-0.5$—values which would indicate that the permutants and wild-type are not from identical folds, yet are more than just randomly similar (Supp. Fig. 7B).

These findings show that the DeepUrfold model is well suited to our task because our encoding is agnostic to topological 'connectivity' information and rather is only sensitive 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 secondary structures are contiguous in sequence–space. This approach is useful in finding similarities amongst sets of seemingly dissimilar 3D structures—and thereby identifying specific candidate urfolds—because two sub-domain portions from otherwise rather (structurally) different domains may be quite similar to each other, even if the domains they are a part of have different (domain-level) topologies but identical overall architectures. This concept can be represented symbolically: for a subset of SSEs, d, drawn from a full domain $\mathcal{D}$, the Urfold model permits relations (denoted by the ' ' symbol) to be detected between two different 'folds', i and j (i.e. $d_i \sim d_j$), without requiring that relation to 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 $d_j \subset \mathcal{D}_j$. Here, we can view the characteristic stringency or 'threshold' level of the urfold 'd' as being 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 the continuous nature of fold space

The latent space of each superfamily (SF)-level DeepUrfold model provides a uniquely informative view of fold space. Each SF model captures the different 3D geometries and physicochemical properties that characterize that individual SF as a single 'compressed' data point; in this way, the latent 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 each SF—is able to reconcile the discrete/continuous dichotomy 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 between two otherwise seemingly unrelated domains (e.g., $d^{\mathrm{SH3}} \sim d^{\mathrm{OB}}$ is not forbidden, using the terminology introduced above).

We represent and analyze the latent space of representative domains for 20 SFs, mapped into two dimensions. Proteins that share similar geometries and biophysical properties have similar embeddings that are close together in this latent-space representation, regardless of the annotated 'true' superfamily. Though this picture of the protein universe is limited to 20 highly populated

CATH SFs (in the present work), already we can see that these SF domains appear to be ordered by secondary structure composition, consistent with past analyses that used approaches such as multidimensional scaling (Hou *et al.*, 2005). Intriguingly, variable degrees of intermixing between SFs can be seen in UMAP projections such as illustrated in Fig. 3. In addition to this mixing, the latent space projection is not punctate: rather, it is fairly 'compact' (in a loose mathematical sense) and well-connected, with only a few disjoint 'outlier' regions. During manual inspection of outlier domain structures, we find that many of them are incomplete sub-domains or a single part of a domain swapped region. 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.), the distributions that these VAE-based SF models each learn (again, as 'good' approximations to the true posterior $p_i(\mathrm{x}_{ij}|\boldsymbol{\theta_i})$) 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 or '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 the latent-space can be seen to follow, as a gradient from 'all-α' domains to 'all-β' domains, separated by 'α/β' domains. This finding is reassuring with respect to previous studies of protein fold space (Hou *et al.*, 2005), as well as the geometric intuition that the similarity between two domains would 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).

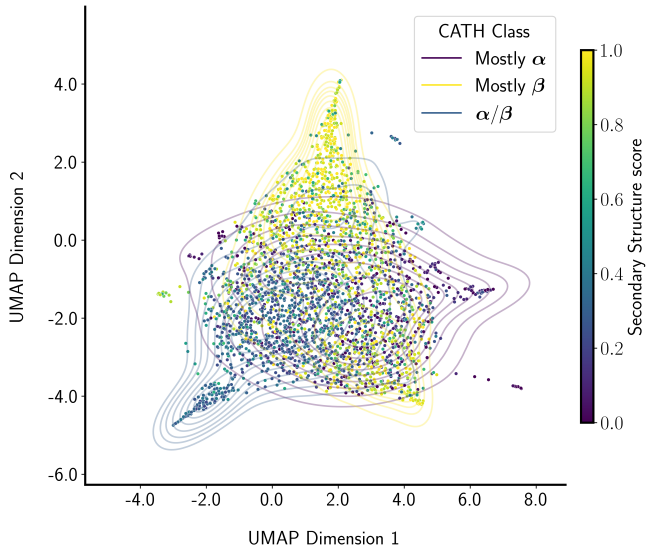

FIG. 3. Dominant variables of the latent space capture gross structural properties and indicate a continuous 'fold space'. As a proof-of-principle, we fit 20 distributions from 20 CATH homologous superfamilies, each being modeled via the DeepUrfold approach. Representatives from each SF were subjected to models trained on domains from the same SF, and the latent space variables for each structural domain were examined via the uniform manifold approximation and projection (UMAP) method to reduce the 1024 dimensions of the actual model to a two-dimensional projection. In this representation, kernel density estimates (contour lines) surround domains with the same annotated CATH *Class*. Each domain is colored by a secondary structure score, showing that they are roughly ordered by secondary structure composition. The secondary structure score is computed as $\frac{\#\beta \text{ atoms} - \#\alpha \text{ atoms}}{2(\#\beta \text{ atoms} + \#\alpha \text{ atoms})} + 0.5$

Protein interrelationships defy discrete clusterings

Our finding that protein fold space is rather continuous implies that there are, on average, webs of interconnections (similarities,

relationships) between a protein $\mathcal{A}$ and its neighbors in fold space ($\mathcal{A}'$, $\mathcal{A}''$, $\mathcal{B}$,...). Therefore, we believe that an optimally realistic view of fold space will not entail hierarchically clustering proteins into mutually exclusive bins. Alternatives to discrete clustering would be *fuzzy clustering*, *multi-label classification*, or *mixed-membership community detection* algorithms. In DeepUrfold, we formulate this labeling/classification problem by fitting an edge-weighted (Peixoto, 2018), mixed-membership (Peixoto, 2015, 2021), hierarchical (Peixoto, 2014) stochastic block model (SBM; (Peixoto, 2017)) to a fully connected bipartite graph that is built from the similarity scores between (i) the VAE-based SF-level models (one part of the bipartite graph), and (ii) representative structural domains from the representative SFs (the other part of the bipartite graph). In our case, we weight each edge by -log(ELBO). Such a bipartite graph can be represented as an adjacency matrix $\mathbf{A}_{d\times sfam}$ and covariate edge weights $\boldsymbol{x}$ (between vertices in the two 'parts' of the bipartite graph), where $sfam \in 20$ representative SFs and $d \in 3654$ representative domains from 20 representative SFs. The likelihood of such a bipartite graph/network occurring by chance—with the same nodes connected by the same edges—is defined by:

$$P(\boldsymbol{A},\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{b},\boldsymbol{e}) = \int P(\boldsymbol{x}|\boldsymbol{A},\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{b},\boldsymbol{e})P(\boldsymbol{e})P(\boldsymbol{\theta})P(\boldsymbol{\gamma})\ \delta\boldsymbol{e}\delta\boldsymbol{\theta}\delta\boldsymbol{\gamma} \times \int P(\boldsymbol{A}|\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{b},\boldsymbol{e})P(\boldsymbol{e})P(\boldsymbol{\theta})P(\boldsymbol{\gamma})\ \delta\boldsymbol{e}\delta\boldsymbol{\theta}\delta\boldsymbol{\gamma}, \quad (2)$$

with $\boldsymbol{\theta}$ as the SBM's latent parameters, $\boldsymbol{x}$ are edge covariate parameters, $\boldsymbol{b}$ represents the blocks (protein communities) in terms of the number of blocks and their membership (which nodes map to which blocks), and $\boldsymbol{e}$ edges may exist between blocks to account for mixed-membership.

The parameters for a given SBM are found using 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 (Peixoto, 2015, 2021).

DeepUrfold's overall methodological approach can be summarized as (i) dataset construction, e.g. via the aforementioned discretization of the 3D structures and biophysical properties into voxelized representations (Draizen et al., in prep); (ii) training of SF-specific models, using 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 model of another SF, $j(\neq i)$; (iv) to detect any patterns amongst these scores, utilization of SBM-based analysis of 'community structure' among the full set of score similarities from the VAE-based SF-level models.

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

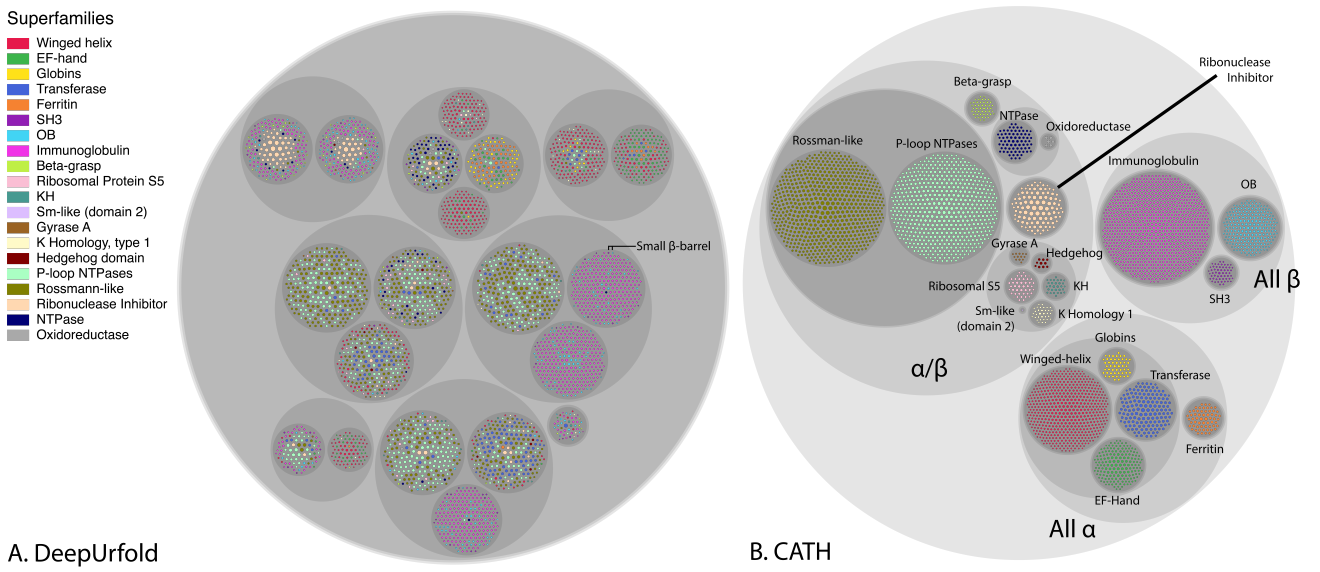

FIG. 4. Protein interrelationships defy discrete clusterings: Stochastic block modeling of an all-vs-all comparison of domain structures and superfamily models. A) We represent the the SBM communities predicted by DeepUrfold as a circle packing diagram following the same hierarchy. Each domain is displayed as the inner most circles (leafs) colored by the annotated CATH superfamily and sized by their number of atoms. All of the superfamily labelled nodes clustered together and were removed from this list (See supplemental file 2). As proof of concept, we show the SH3 and OB domains are found within the same communities. B) CATH Hierarchy represented as a circle packing diagram showing that DeepUrfold is learning a completely different hierarchy.

models, in an exhaustive $all_{\text{SF-models}} \times all_{\text{SF-reps}}$ analysis, reveals the overall 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 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. Clustering 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 structure/function properties.

Domains that have similar ELBO scores against different superfamily models are more likely to contain important biophysical properties at particular (and, presumably, functionally important) locations in 3D space for that superfamily. Furthermore, if two domains are in the same SBM community, it is likely that both domains share the same scores when run through each superfamily (i.e. an inference calculation), so we hypothesize it might contain an urfold that subsumes those two domains (agnostic of whatever SFs they are labeled as belonging to in CATH or other databases). We also expect domains to be in multiple communities, which may represent a protein being constructed of

several 'urfold' or sub-domain elements. However, due to the complexities of analyzing such high dimensional data, we only show the most likely cluster each domain belongs to.

Because our model uses a different input representation of proteins that intentionally ignores all topological/connectivity information, we expect that our model will be least similar to CATH in terms of SBM-related measures such as partition overlap, homogeneity, and completeness (Peixoto, 2021).

Due to the stochastic nature of the SBM, we ran 6 different replicas. While each replica produced slightly different hierarchies and number of clusters (19-23), the communities at lowest level remained consistent with varying degrees of intermixing. In each of the replicas, SH3 and OB clustered into the same communities as well as Rossman-like and P-loop NTPases, instead of their own individual clusters—consistent with the Urfold view of these particular SFs, as predicted based on manual/visual analysis (Mura *et al.*, 2019). In Fig. 4, we chose to display the replica with 20 superfamilies and highest overlap score compared to CATH in order to enable easy comparison with CATH. Most notably, each community contains domains from different superfamilies, consistent with the Urfold model (Fig. 4A). 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) and SH3 and OB cluster together in the same communities (Fig. 4A).

In addition to coloring each domain node by CATH superfamily in the circle-packing diagrams, we also explored coloring domain nodes by secondary structure, average electrostatic potential, average partial charge, and enriched GO terms (Supp. Fig. 12-17; `https://bournelab.org/research/DeepUrfold/`). Interestingly, domains with similar average electrostatic potentials (Supp. Fig. 12) and partial charges (Supp. Fig. 13) can be found to cluster into similar groups, whereas the CATH circle-packing diagrams colored by those same features have no discernable order or structuring; whether or not this phenomenon stems from any underlying, functionally-relevant 'signal' is a question of interest in further work.

In order to assess how 'well' our DeepUrfold model does, we compare our clustering results to 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 compare our DeepUrfold results to a well-established system (e.g., CATH) with the awareness that these are fundamentally different approaches to representing and describing the protein universe. Given all this, models that differ from CATH—versus matching or recapitulating it—can be considered as

representing an alternative view of the protein universe. Somewhat counterintuitively, we deem poorer comparison metrics (e.g., less similarity to CATH) as providing stronger support for the Urfold model of protein structure. Simultaneously, we compare how well other, independently-developed sequence- and structure-based models can reconstruct CATH (Fig. 5). Among all these methods, our DeepUrfold approach produces results 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 show that many other algorithms have difficulty reconstructing CATH, possibly due to the extensive manual curation of CATH, but much more closely reproduce CATH than does our method—we suspect that this is due, in large part, to DeepUrfold's incorporation and integration of more *types* of information than purely 3D geometry.

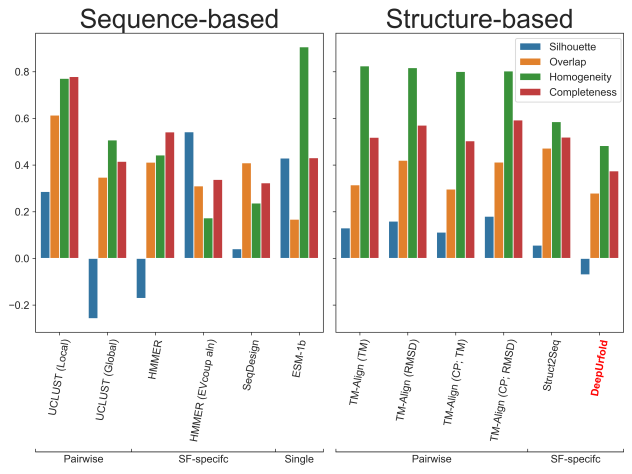

FIG. 5. DeepUrfold does not recapitulate CATH. We compare DeepUrfold to other sequence- and structure-based protein similarity tools by attempting to reconstruct CATH. The scores from each of the algorithms are used as edge weights in the SBM. If scores were increasing e.g. were a distance metric, the converted to a similarity metric by -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 metric of Silhouette Score, overlap, homogeneity, and completeness, a value of 1 is deemed best. DeepUrfold does poorly based for each metric because it does not produce the same clusters, and is learning something completely different compared to the other algorithms. For TM-Align, 'CP' stands for Circular Permutation. For more information, see Supp Table 2.

Discussion

This work has presented a new deep learning-based approach, termed 'DeepUrfold', aimed at systematically identifying putative new urfolds. Notably, the DeepUrfold framework (i) 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, most notably, is alignment-free (as it leverages latent-space embeddings of structure, versus direct 3D 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 they can be encoded 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 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 (e.g., using 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 sequence/structure/function interrelationships. This is intentional: molecular evolution acts on the sequence/structure/function triad as its base 'entity', not on purely geometric/3D structure alone.

We demonstrate (i) the general utility of this new type of similarity metric for representing and comparing protein domain structures, based on deep generative models, and (ii) that a mixed-membership community detection algorithm can identify what we previously found (via manual/visual analysis) to be putative urfolds. Finally, we emphasize that because DeepUrfold is agnostic of precise protein topology (i.e., order of SSEs in 3-space), higher levels of similarity can be readily detected ('higher' than CATH's 'T' level, below its 'A' level), including the potential of non-contiguous fragments. We believe that such such spatially-compact groups of frequently recurring sub-domain 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 (Skolnick *et al.*, 2009).

Overall, the DeepUrfold framework provides a sensitive approach to detect and thus explore distant protein inter-relationships, 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 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 protein embeddings is ill-advised 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 proteins, as an urfold) in this fold space is reflective of underlying evolutionary processes, e.g. akin to Edwards & Deane's finding that "evolutionary information is encoded along these structural bridges [in fold space]" (Edwards and Deane, 2015)).

An informative next step would be to use DeepUrfold to identify structural fragments that contain similar patterns of geometry and biophysical properties between proteins from very different superfamilies. Notably, these fragments may be continuous or discontinuous, and pursuing this goal might help unify the 'primordial peptides' (Alva *et al.*, 2015) and 'themes' (Nepomnyachiy *et al.*, 2017) concepts with the Urfold hypothesis, allowing

connections between unexplored (or at least under-explored) regions of fold space. We suspect that 'Explainable AI' techniques, such as Layer-wise Relevance Propagation (LRP; (Hochuli *et al.*, 2018; Montavon *et al.*, 2019)), 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 (i.e., 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: thus, we can probe the neural network 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, physicochemical 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. Such communities—beyond helping to detect and define new urfolds—may offer a novel perspective on remote protein homology.

Materials and Methods

Dataset

We create the 'Prop3D' dataset using the 20 CATH superfamilies of interest (Fig. 1C; Supp. Table 1). Domain structures from each of the 20 superfamilies are 'cleaned' by adding missing residues with MODELLER (Eswar *et al.*, 2006), missing atoms with SCWRL4 (Krivov *et al.*, 2009), and protonating and energy minimizing (simple de-bump) with PDB2PQR (Dolinsky *et al.*, 2007). Next, we compute a host of derived properties for each domain in CATH (Draizen et al., in prep)–including (i) include purely geometric/structural quantities, e.g. secondary structure (Kabsch and Sander, 1983), solvent accessibility, (ii) physicochemical properties, e.g. hydrophobicity, partial charges, electrostatic potentials, and (iii) basic chemical descriptors (atom and residue types). The computation was performed using the Toil workflow engine (Vivian *et al.*, 2017) and data was stored using the Hierarchical Data Format (version 5) in the Highly Scalable Data Service (HSDS). The domains from each superfamily were split such that all members of a S35 35% sequence identity cluster (pre-calculated by CATH) were on the same side of the split. We split them roughly 80% training, 10% validation, and 10% test (Draizen et al., in prep; https://doi.org/10.5281/zenodo.6873024).

Each atom was attributed with 7 groups of boolean (one-hot encoded) features: (1)

Atom Type (C,CA,N,O,OH,Unknown); (2) Residue Type* (ALA, CYS, ASP, GLU, PHE, GLY, HIS, ILE, LYS, LEU, MET, ASN, PRO, GLN, ARG, SER, THR, VAL, TRP, TYR, Unknown); (3) Secondary Structure (Helix,Sheet,Loop/Unknown); (4) Hydrophobic; (5)Electronegative; (6) Positively Charged; and (7) Not exposed to solvent. For all models reported, residue type was removed because it was was found to be uninformative for this type of representation (Supp. Fig. 3).

Protein Representation

We represent protein domains as voxels, or 3D volumetric pixels. Briefly, our method centers protein domains in a 256^3 $Å^3$ cube volume to allow large domains, and each atom is mapped to $1Å^3$ voxels using a kD-tree data structure with a query ball radius set to the van der Waals radius of the atom. If two atoms share the space in a given voxel, the maximum between their feature vectors is used because they all contain binary values. Because a significant fraction of voxels do not contain any atoms, protein domain structures can be encoded via a sparse representation; this substantially mitigates the computational costs of our deep learning workflow using MinkowskiEngine (Choy *et al.*, 2019).

Because there is no 'correct' orientation of a protein domain, we applied random rotations to each protein domain structure; these rotations were in the form of orthogonal transformation matrices drawn from the Haar distribution, which is the uniform distribution on the 3D rotation group (i.e. SO(3); (Stewart, 1980)).

VAE Model Design and Training

A sparse 3D-CNN variational autoencoder was adapted from MinkowskiEngine (Choy *et al.*, 2019; Gwak *et al.*, 2020). In the Encoder, there are 7 blocks consisting of Convolution (n->2n), BatchNorm, ELU, Convolution (2n->2n), BatchNorm, and ELU, where n=[16,32,64,128,256,512,1024], doubling at each block. Finally, the tensors are pooled using Global pooling, and the model outputs both a normal distribution's mean and log variance. Next, the learned distribution is sampled from and used as input into the Decoder. In the decoder, there are also 7 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 $264^3 Å^3$ volume.

In VAEs, a 'reparameterization trick' allows for backpropagation through random variables 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 $\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \mathcal{N}(0,\mathbf{I})$, where $\odot$ denotes the Hadamard (element-wise) matrix product and $\mathcal{N}$ is the 'auxiliary noise' term (Kingma and Welling, 2013).

We optimize against the Evidence Lower BOund (ELBO) described in equation 1, which combines (a) the mean squared error (MSE) of the reconstructed domain and (b) the difference between the learned distribution and the true distribution of the SF (i.e., the Kullback–Leibler [KL] divergence, or relative entropy) (Kingma and Welling, 2013).

We used stochastic gradient descent (SGD) as the optimization algorithm, 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 around 110M parameters in total and all the networks were trained for 30 epochs, using a batch size of 255. We used the open-source frameworks PyTorch (Paszke *et al.*, 2019) and PytorchLightning (Falcon *et al.*, 2020) to simplify training and inference and make the models more reproducible.

In order to determine the best hyperparameters for the VAE, we used Weights & Biases Sweeps (Biewald, 2020) to search over the batch size, learning rate, convolution kernel size, transpose convolution kernel size, and convolution stride in the Ig model while optimizing the ELBO. We used the Bayesian Optimization search strategy and hyperband method with 3 iterations for early termination. We found no significant changes and used the 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, we follow the One-Class Classifier approach, creating one VAE for each superfamily. We also train a joint SH3 and OB model and compare random over- and under-sampling from ImbalancedLearn (Lemaître *et al.*, 2017) on joint models of multiple superfamilies (Supp. Fig. 8).

Evaluation of Model Performance

We calculate the area under the Receiver Operating Characteristic curve (auAUC) and the area under the precision-recall curve (auPRC) for 20 SFs. Representative domains, as defined by CATH, for each superfamily were subjected to their SF-specific VAE and predicted values were micro-averaged to perform auROC and auPRC calculations. Immunoglobulins were chosen to display in the supplemental material for this paper (Supp. Fig. 4-6), but the results for all SFs can be found in the extended supplemental material. All SFs report similar metrics for each group of features.

Assess the Urfold model by subjecting proteins with permuted secondary structures to the superfamily-specific VAEs

To gauge the sensitivity of our DeepUrfold model to loop orderings (topology), we generate fictitious folds by implementing a multi-loop permutation algorithm (Dai and Zhou, 2011) in order to 'scramble' the secondary structural

elements (SSEs) found in a representative SH3 and OB domains. We stitch together the SSEs and relax new 3D structures using MODELLER (Eswar *et al.*, 2006).

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. We also calculate a 'background' distribution of each model by perming an all vs all TM-align for all domains in our representative CATH domains, saving domain that have a TM-Score ≤ 0.3 as that is thought to represent domains that have random similarity.

Latent-space Organization

We subject representative domains from a single superfamily through its superfamily model and visualize the latent space of each representative. A 'latent-space' for a given domain corresponds to a 1024 dimensional vector describing the representatives in their most 'compressed' form, accounting for the position of each atom and their biophysical properties represented by the mean of the learned distribution. We combine the latent spaces from each domain from each superfamily and then reduce the number of dimensions to two in order to easily visualize it; the latter is achieved using the uniform manifold approximation and projection (UMAP) algorithm. UMAP is a dimensionality reduction algorithm that is similar to methods such as PCA (principal component analysis; Supp. Fig. 9) and particularly t-SNE (t-distributed stochastic neighbor embedding; Supp. Fig. 10), but preserves topological relationships at both local and global scales in a dataset.

Mixed-membership Community Detection

We performed all-vs-all comparisons of domains and superfamilies by subjecting representative protein domain structures from each of the 20 chosen SF through each SF-specific one-class VAE model. The ELBO loss score for each domain—SF-model pair can be used to quantitatively evaluate pairwise 'distances' between SFs by treating it as a fully connected bipartite graph between domains and SF models, with edges weighted by the -log(ELBO) score. Stochastic Block Models (SBM; (Peixoto, 2017)) are a generative model for random graphs that can be used to partition the bipartite graph into communities of domains that have similar distribution of edge covariates between them (Peixoto, 2018).

Using the same SBM approach as we did for DeepUrfold, we compare our results to state-of-the-art sequence- and structure-based methods for comparing proteins. All SBMs are created using fully connected bipartite graphs connected n CATH S35 domains to m Superfamily models. In this case, we used 3654 representative CATH domains from 20 superfamilies, creating a 3654

$\times$ 20 similarity matrix for each method we wish to compare. Each SBM was degree corrected, overlapping, and nested and fit to a real normal distribution of edge covariates. For methods with decreasing scores (closer to zero is best), we took the negative log of each score, whereas scores from methods with increasing scores remained the same.

While only the 'Superfamily-specific' models are directly comparable (e.g. where $n \times m$ matrices are the original output created by subjecting n CATH representative domains without labels to m superfamily-specific models), we also included 'Pairwise' and 'Single Model' methods. For pairwise approaches, an all-vs-all $n \times n$ similarity matrix is created and is converted to an $n \times m$ by taking the median distance of a single CATH domain to every other domain in a given superfamily. 'Single Model' approaches are where 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 it into an $n \times m$ matrix, we take the median distance of a single CATH domain embedding to every other domain embedding from a given superfamily.

Because we have no ground truth with the Urfold view of the protein universe, we perform cluster comparison metrics on each SBM community compared to the original CATH clusterings; these measures can include partition overlap, homogeneity, and completeness for each of the protein comparison tools:

- **Silhouette Score:** measure of how similar an object is to its own cluster (cohesion) compared to next closest cluster (separation). -1: incorrect, 0: perfect, 1: too dense
- **Overlap:** maximum overlap between partitions by solving an instance of the maximum weighted bipartite matching problem (Peixoto, 2021)
- **Homogeneity:** each cluster contains only members of a single class. [0, 1], 1=best
- **Completeness:** all members of a given class are assigned to the same cluster. [0, 1], 1=best

All comparisons start using the sequence and structure representatives from CATH's S35 cluster for each of the 20 superfamilies of interest. USEARCH (Edgar, 2010) was run twice with parameters `-allpairs_local` and `-allpairs_global`; both runs included the `-acceptall` parameter. HMMER (Mistry *et al.*, 2013) models were built using (1) MUSCLE (Edgar, 2004) alignments from CATH's S35 cluster; and (2) a deep MSA created from EVcouplings (Hopf *et al.*, 2019) using jackhmmer (Mistry *et al.*, 2013) 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 (Shin *et al.*, 2021) was run using the same MSAs from EVcouplings. We also compared against the pretrained ESM models (Rives *et al.*, 2021).

For other structure-based comparisons, we ran TM-Align (Zhang and Skolnick, 2005) on all representative domains with and without circular permutations saving RMSD and TM-Scores. Struct2Seq (Ingraham *et al.*, 2019) was run with default parameters after converting domain structure representatives into dictionaries matching the required input.

Data Availability

The Prop3D dataset used to train each superfamily model can found at `https://doi.org/10.5281/zenodo.6873024`, which includes the raw HDF file as well as instructions to access the public version of the dataset on the University of Virginia Research Computing HSDS endpoint `http://hsds.uvarc.io` (in prep).

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`.

All code to build datasets *and* train models can be found at `http://github.com/bouralab/Prop3D` and `http://github.com/bouralab/DeepUrfold` respectively.

We also provide an accompanying website to explore the SBM communities and the CATH hierarchy at `https://bournelab.org/research/DeepUrfold/`

Acknowledgements

We thank Luis Felipe Murillo and John Readey for support with HSDS, as well as Jaime Iranzo and Tiago Peixoto for the idea and support in using stochastic block models. Portions of this work were supported by the University of Virginia and NSF Career award MCB-1350957. ED was supported by the University of Virginia Presidential Fellowship in Data Science.

References

Agrawal, V. and Kishan, R. K. 2001. Functional evolution of two subtly different (similar) folds. *BMC structural biology*, 1(1): 1–6.

Alva, V., Söding, J., and Lupas, A. N. 2015. A vocabulary of ancient peptides at the origin of folded proteins. *eLife*, 4: e09410.

Alvarez-Carreño, C., Gupta, R. J., Petrov, A. S., and Williams, L. D. 2022. The evolution of protein folds by creative destruction. *bioRxiv*.

Alvarez-Carreño, C., Penev, P. I., Petrov, A. S., and Williams, L. D. 2021. Fold evolution before LUCA: Common ancestry of SH3 domains and OB domains. *Molecular Biology and Evolution*, 38(11): 5134–5143.

Alvarez-Carreño, C., Gupta, R., Petrov, A. S., and Williams, L. D. 2022. Protein fold evolution by creative destruction. *BioRxiv*.

Andreeva, A., Howorth, D., Chothia, C., Kulesha, E., and Murzin, A. G. 2014. SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Research*, 42(Database issue): D310–4.

Biewald, L. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

Bromberg, Y., Aptekmann, A. A., Mahlich, Y., Cook, L., Senn, S., Miller, M., Nanda, V., Ferreiro, D. U., and Falkowski, P. G. 2022. Quantifying structural relationships of metal-binding sites suggests origins of

biological electron transfer. *Sci. Adv.*, 8(2).

Budowski-Tal, I., Nov, Y., and Kolodny, R. 2010. 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(8): 3481–3486.

Cheng, H., Schaeffer, R. D., Liao, Y., Kinch, L. N., Pei, J., Shi, S., Kim, B.-H., and Grishin, N. V. 2014. ECOD: an evolutionary classification of protein domains. *PLoS Computational Biology*, 10(12): e1003926.

Choy, C., Gwak, J., and Savarese, S. 2019. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3075–3084.

Dai, L. and Zhou, Y. 2011. Characterizing the existing and potential structural space of proteins by large-scale multiple loop permutations. *Journal of Molecular Biology*, 408(3): 585–595.

Dolinsky, T. J., Czodrowski, P., Li, H., Nielsen, J. E., Jensen, J. H., Klebe, G., and Baker, N. A. 2007. PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations. *Nucleic Acids Research*, 35(Web Server issue): W522–5.

Durairaj, J., Akdel, M., de Ridder, D., and van Dijk, A. D. J. 2020. Geometricus represents protein structures as shape-mers derived from moment invariants. *Bioinformatics*, 36(Suppl_2): i718–i725.

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5): 1792–1797.

Edgar, R. C. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, 26(19): 2460–2461.

Edwards, H. and Deane, C. M. 2015. Structural bridges through fold space. *PLoS Computational Biology*, 11(9): e1004466.

Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M.-Y., Pieper, U., and Sali, A. 2006. Comparative protein structure modeling using modeller. *Current Protocols in Bioinformatics*, Chapter 5: Unit 5.6.

Falcon, W., Borovec, J., Wälchli, A., Eggert, N., Schock, J., Jordan, J., Skafte, N., Ir1dXD, Bereznyuk, V., Harris, E., Murrell, T., Yu, P., Præsius, S., Addair, T., Zhong, J., Lipin, D., Uchida, S., Bapat, S., Schröter, H., Dayma, B., Karnachev, A., Kulkarni, A., Komatsu, S., Martin.B, SCHIRATTI, J.-B., Mary, H., Byrne, D., Eyzaguirre, C., cinjon, and Bakhtin, A. 2020. Pytorchlightning/pytorch-lightning: 0.7.6 release.

Fox, N. K., Brenner, S. E., and Chandonia, J.-M. 2014. SCOPe: Structural classification of proteins–extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Research*, 42(Database issue): D304–9.

Friedberg, I. and Godzik, A. 2005. Fragnostic: walking through protein structure space. *Nucleic Acids Research*, 33(Web Server issue): W249–51.

Goncearenco, A., Shaytan, A. K., Shoemaker, B. A., and Panchenko, A. R. 2015. Structural perspectives on the evolutionary expansion of unique protein-protein binding sites. *Biophysical Journal*, 109(6): 1295–1306.

Grishin, N. V. 2001. Fold change in evolution of protein structures. *Journal of Structural Biology*, 134(2-3): 167–185.

Gwak, J., Choy, C. B., and Savarese, S. 2020. Generative sparse detection networks for 3d single-shot object detection. In *European conference on computer vision*.

Harrison, A., Pearl, F., Mott, R., Thornton, J., and Orengo, C. 2002. Quantifying the similarities within fold space. *Journal of Molecular Biology*, 323(5): 909–926.

Hochuli, J., Helbling, A., Skaist, T., Ragoza, M., and Koes, D. R. 2018. Visualizing convolutional neural network protein-ligand scoring. *Journal of molecular graphics & modelling*, 84: 96–108.

Holm, L. and Sander, C. 1996. Mapping the protein universe. *Science*, 273(5275): 595–603.

Hopf, T. A., Green, A. G., Schubert, B., Mersmann, S., Schärfe, C. P. I., Ingraham, J. B., Toth-Petroczy, A., Brock, K., Riesselman, A. J., Palmedo, P., Kang, C., Sheridan, R., Draizen, E. J., Dallago, C., Sander, C., and Marks, D. S. 2019. The EVcouplings python framework for coevolutionary sequence analysis. *Bioinformatics*, 35(9): 1582–1584.

Hou, J., Jun, S.-R., Zhang, C., and Kim, S.-H. 2005. 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(10): 3651–3656.

Ingraham, J., Garg, V., Barzilay, R., and Jaakkola, T. 2019. Generative models for graph-based protein design. *Advances in Neural Information Processing Systems*.

Kabsch, W. and Sander, C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12): 2577–2637.

Kinch, L. N. and Grishin, N. V. 2002. Evolution of protein structures and functions. *Current Opinion in Structural Biology*, 12(3): 400–408.

Kingma, D. P. and Welling, M. 2013. Auto-encoding variational bayes. *arXiv*.

Kolodny, R., Koehl, P., Guibas, L., and Levitt, M. 2002. Small libraries of protein fragments model native protein structures accurately. *Journal of Molecular Biology*, 323(2): 297–307.

Kolodny, R., Pereyaslavets, L., Samson, A. O., and Levitt, M. 2013. On the universe of protein folds. *Annual review of biophysics*, 42: 559–582.

Kolodny, R., Nepomnyachiy, S., Tawfik, D. S., and Ben-Tal, N. 2021. Bridging themes: short protein segments found in different architectures. *Molecular Biology and Evolution*, 38(6): 2191–2208.

Krishna, S. S. and Grishin, N. V. 2005. Structural drift: a possible path to protein fold change. *Bioinformatics*, 21(8): 1308–1310.

Krivov, G. G., Shapovalov, M. V., and Dunbrack, R. L. 2009. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins*, 77(4): 778–795.

Lemaître, G., Nogueira, F., and Aridas, C. K. 2017. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17): 1–5.

Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A., and Punta, M. 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Research*, 41(12): e121.

Montavon, G., Binder, A., Lapuschkin, S., Samek, W., and Müller, K.-R. 2019. Layer-wise relevance propagation: An overview. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, editors, *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700 of *Lecture notes in computer science*, pages 193–209. Springer International Publishing, Cham.

Mura, C., Randolph, P. S., Patterson, J., and Cozen, A. E. 2013. Archaeal and eukaryotic homologs of hfq. *RNA Biology*, 10(4): 636–651. PMID: 23579284.

Mura, C., Veretnik, S., and Bourne, P. E. 2019. The urfold: Structural similarity just above the superfold level? *Protein Science*, 28(12): 2119–2126.

Nepomnyachiy, S., Ben-Tal, N., and Kolodny, R. 2017. 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(44): 11703–11708.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. 2019. Pytorch: An imperative style, high-performance deep learning

library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Peixoto, T. P. 2014. Hierarchical block structures and high-resolution model selection in large networks. *Physical Review X*, 4(1).

Peixoto, T. P. 2015. Model selection and hypothesis testing for large-scale network models with overlapping groups. *Physical Review X*, 5(1).

Peixoto, T. P. 2017. Nonparametric bayesian inference of the microcanonical stochastic block model. *Physical Review E*, 95(1).

Peixoto, T. P. 2018. Nonparametric weighted stochastic block models. *Physical review. E*, 97(1-1): 012306.

Peixoto, T. P. 2021. Revealing consensus and dissensus between network partitions. *Physical Review X*, 11(2): 021003.

Prati, R. C., Batista, G. E. A. P. A., and Monard, M. C. 2009. Data mining with imbalanced class distributions: concepts and methods.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., and Fergus, R. 2021. 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(15).

Sadreyev, R. I., Kim, B.-H., and Grishin, N. V. 2009. Discrete-continuous duality of protein structure space. *Current Opinion in Structural Biology*, 19(3): 321–328.

Shin, J.-E., Riesselman, A. J., Kollasch, A. W., McMahon, C., Simon, E., Sander, C., Manglik, A., Kruse, A. C., and Marks, D. S. 2021. Protein design and variant prediction using autoregressive generative models. *Nature Communications*, 12(1): 2403.

Sillitoe, I., Dawson, N., Lewis, T. E., Das, S., Lees, J. G., Ashford, P., Tolulope, A., Scholes, H. M., Senatorov, I., Bujan, A., Ceballos Rodriguez-Conde, F., Dowling, B., Thornton, J., and Orengo, C. A. 2019. CATH: expanding the horizons of structure-based functional annotations for genome sequences. *Nucleic Acids Research*, 47(D1): D280–D284.

Skolnick, J., Arakaki, A. K., Lee, S. Y., and Brylinski, M. 2009. 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(37): 15690–15695.

Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S. J., and Söding, J. 2019. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics*, 20(1): 473.

Stewart, G. W. 1980. The efficient generation of random orthogonal matrices with an application to condition estimators. *SIAM Journal on Numerical Analysis*, 17(3): 403–409.

Taylor, W. R. 2020. Exploring protein fold space. *Biomolecules*, 10(2).

Vivian, J., Rao, A. A., Nothaft, F. A., Ketchum, C., Armstrong, J., Novak, A., Pfeil, J., Narkizian, J., Deran, A. D., Musselman-Brown, A., Schmidt, H., Amstutz, P., Craft, B., Goldman, M., Rosenbloom, K., Cline, M., O'Connor, B., Hanna, M., Birger, C., Kent, W. J., Patterson, D. A., Joseph, A. D., Zhu, J., Zaranek, S., Getz, G., Haussler, D., and Paten, B. 2017. Toil enables reproducible, open source, big biomedical data analyses. *Nature Biotechnology*, 35(4): 314–316.

Xu, J. and Zhang, Y. 2010. How significant is a protein structure similarity with tm-score = 0.5? *Bioinformatics*, 26(7): 889–895.

Youkharibache, P. 2019. Protodomains: Symmetry-related supersecondary structures in proteins and self-complementarity. *Methods in Molecular Biology*, 1958: 187–219.

Youkharibache, P., Veretnik, S., Li, Q., Stanek, K. A., Mura, C., and Bourne, P. E. 2019. The small β-barrel domain: A survey-based structural analysis. *Structure*, 27(1): 6–26.

Zhang, Y. and Skolnick, J. 2005. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, 33(7): 2302–2309.