A Structure-Aware Generative Framework for Exploring Protein Sequence and Function Space

Divyanshu Shukla[3], Jonathan Martin[4], Faruck Morcos[4,5,6], Davit A. Potoyan[1,2,3,*]

1. Department of Chemistry, Iowa State University, Ames, IA 50011
2. Roy J. Carver Department of Biochemistry, Biophysics and Molecular Biology
3. Bioinformatics and Computational Biology Program, Iowa State University
4. Department of Biological Sciences, UT Dallas, TX 75080
5. Departments of Bioengineering and Physics, UT Dallas, TX 75080
6. Center for Systems Biology, UT Dallas, TX 75080

*correspondence: potoyan@iastate.edu

Abstract

The rapid expansion of protein sequence databases has far outpaced experimental structure determination, leaving many unannotated sequences, particularly the more remote homologs with low sequence identity. Because protein folds are more conserved and functionally informative than sequences alone, structural information offers a powerful lens for analysis. Here, we introduce a generative, structure-aware framework that integrates geometric encoding and coevolutionary constraints to map, cluster, and design protein sequences. Our approach employs the 3D interaction (3Di) alphabet to convert local residue geometries into compact, 20-state discrete representations. Using ProstT5, we enable bidirectional translation between amino acid sequences and 3Di representations, facilitating sensitive homology detection and structure-guided sequence generation. We construct a latent sequence landscape by combining 3Di-based alignments with direct coupling analysis (DCA) and variational autoencoders (VAE), unifying tasks such as clustering, annotation, and design. This integrative framework enhances the detection of coevolutionary signals and enables rational sampling of structural variants, even without functional labels. We demonstrate the utility of our method across diverse protein families, including globins, kinases, and malate dehydrogenases, achieving improved contact prediction, homology inference, and sequence generation. Together, our approach offers a quantitative, generative view of protein structure space, advancing protein evolution, and design studies.

Introduction

Proteins underpin a vast array of cellular functions, from energy metabolism to cell division. Deciphering their three-dimensional (3D) structures is key to understanding function, tracing evolutionary relationships, and guiding drug design[1–3]. While protein sequence databases now contain hundreds of millions of entries, experimentally determined structures remain limited due to the time and cost of traditional methods. Recent advances in computational prediction—most notably AlphaFold2 have revolutionized structural biology by providing high-accuracy models at scale[4,5]. These models now support diverse applications, including structural alignment, pocket detection, complex modeling, novel fold discovery, and genome annotation refinement[6–8]. Yet despite the success of sequence-based tools, remote homology detection remains a major bottleneck, leaving a substantial fraction of proteins functionally unannotated[9,10]. Because protein folds are more conserved than sequences, especially below 30% identity, structure-based analysis provides a powerful alternative for identifying distant homologs and revealing functional and evolutionary relationships[11,12].

The 3D interaction (3Di) alphabet encodes the spatial relationship between each residue and its nearest neighbor into one of 20 discrete geometric states, enabling scalable structure-based comparisons. Compared to classical structural alphabets, 3Di offers lower sequential dependency, more balanced state distributions, and higher information density localized within conserved structural cores[1,13]. ProstT5 builds upon this by fine-tuning the ProtT5 language model to translate bidirectionally between amino acid and 3Di sequences[14]. This enables sensitive remote homology detection, rapid structure-aware searches, and de novo sequence generation from structural input[15,16]. Here, we extend this capability by introducing a generative latent space that integrates 3Di representations with coevolutionary constraints from direct coupling analysis (DCA). Using a variational autoencoder (VAE), we construct a structurally informed sequence landscape that unifies clustering, annotation, and design [17,18]. By capturing meaningful structural variation while suppressing noise, the 3Di encoding enhances DCA signal detection and improves organization within latent space[19].

Our dual-representation framework integrates the evolutionary constraints of direct coupling analysis (DCA) with the generative capacity of variational autoencoders (VAE), enabling the rational sampling of structural variants across diverse protein families without requiring functional labels. We demonstrate its utility across several families, showcasing improvements in structural contact prediction, functional feature identification, remote homology detection, phylogenetic relationships and guided sequence generation via the ProstT5 decoder. Together, these advances yield a new quantitative and conceptual map of protein structure space, offering insights into fitness landscapes, evolutionary diversification, and protein design.

Methods

1. MSA Acquisition and 3Di Sequence Generation

For the analysis, multiple sequence alignments (MSAs) were sourced from various origins. The MSA for malate dehydrogenase was obtained using the HMMSearch against the UniProt database, utilizing the GREMLIN [20,21]. MSAs for globins and TRPM domains were referenced from training dataset of LGL-VAE[18], and kinases MSAs from [22,23]. All alignments were saved in FASTA format. To translate these amino acid sequences into 3Di sequences, we employed ProstT5—a state-of-the-art, pre-trained protein language model—on three Nvidia A100 GPUs. ProstT5, an extension of ProtT5, encodes both sequence and structural information into 3Di tokens. The generated 3Di sequences were then reformatted to align with their respective amino acid MSAs, ensuring consistency for subsequent variational autoencoder analyses[13,15].

2. VAE model architecture

The variational autoencoder (VAE) is employed to generate data samples using a latent variable($x \in X$) model with parameters θ . This model defines a prior distribution over latent variables z, denoted as $p(z)$. The marginal likelihood of an observed data point x under the model is expressed as:

$$p_\theta(x) = \int p_\theta(x|z) dp_\theta(z) \tag{1}$$

However, since both the model parameters θ and the latent variables z are unknown, directly evaluating the marginal likelihood (Equation 1) becomes computationally intractable—particularly as the number of parameters increases. To address this challenge, the approach introduced in [17] proposes approximating the true posterior distribution $p(z \mid x)$ with a separate, tractable distribution $q_\phi(z \mid x)$, parameterized by ϕ:

$$q_\phi(z|x) \approx p_\theta(z|x) \tag{2}$$

The model parameterized by ϕ, which approximates the posterior $q_\phi(z \mid x)$, is referred to as the encoder, while the model parameterized by θ, which defines the likelihood $p_\theta(x \mid z)$, is known as the decoder. Together, these models enable sampling and reconstruction of data through a latent representation.

$$\log p_\theta(x) = D_{KL}\big(q_\phi(z|x) \big|\big| p_\theta(z|x)\big) + \zeta(\theta, \phi, x) \tag{3}$$

Here, D_{KL} denotes the Kullback–Leibler divergence, which measures how well the encoder's approximate posterior $q_\phi(z \mid x)$ aligns with the true posterior defined by the decoder. The second term in the expression represents a variational lower bound on the model's fit to the marginal distribution over z. This leads to a simplified training objective known as the Evidence Lower Bound (ELBO), expressed as:

$$\text{ELBO} = -E_{z\sim q_\theta(Z|X)} \log\big(p_\theta(x|z)\big) + D_{KL}(q_\theta(z|x) \,||\, p_\theta(z)) \tag{4}$$

This expression defines the objective function minimized during VAE training. The first term of the ELBO represents the reconstruction loss, quantifying how accurately the decoder can reconstruct the original input from the latent representation. The second term measures how closely the learned latent distribution $q_\phi(z \mid x)$ aligns with the assumed prior distribution $p(z)$, using the Kullback–Leibler divergence as a regularization. To enable backpropagation through the stochastic sampling process, we apply the reparameterization trick. Specifically, the encoder is designed to output the mean μ and standard deviation σ of a Gaussian distribution. These parameters are then combined with an auxiliary noise variable $\varepsilon \sim N(0, I)$ to produce latent variables z as:

$$z = \mu + \sigma \odot \varepsilon \tag{5}$$

This reparametrized z forms the input the decoder utilizes for sequence generation, allowing us to define $p_\theta(z)$ as a Gaussian distribution, thereby providing an analytical solution to the gradient of Equation (4).

3. Data representation and decoding

In our implementation, 3Di sequences are represented as one-hot encoded vectors. For a sequence of length L, the input array has dimensions 20 × L, where each column corresponds to one position in the sequence. Each column contains a single entry with value 1, indicating the identity of the 3Di state at that position, and 0s elsewhere. The 20 possible entries encode the 20 3Di states. During decoding, the latent variables z are passed through a Softmax activation, producing a probability distribution over the 20 possible 3Di types at each sequence position. Thus, the output layer has the same dimensions as the input: $R^{23\times L}$, where each column represents a probability distribution over sequence symbols at a given residue position.

$$p(a|z)_i = \frac{\exp\left(\psi_{a_i}(z)\right)}{\sum_{k\epsilon A} \exp\left(\psi_{k_i}(z)\right)} \tag{6}$$

This yields L rows with probability values summing to one in each row. The reconstruction error term in Equation (4) evaluates to zero if the input and output matrices are identical, indicating that the only possible sequence at some point z is the input sequence.

4. Hyperparameters and Training

All models were trained using 3 × L hidden units in both the encoder and decoder, where L is the input sequence length. The ReLU activation function was applied to all hidden layers. A latent dimensionality of 2 was used. Optimization was performed using the Adam optimizer with a learning rate of 1×10^{-4}, and L2 regularization with a penalty of 1×10^{-4} was applied to the hidden units. Training was terminated early if the loss failed to improve over 50 consecutive epochs. Empirically, we observed that increasing the number of hidden units beyond 3 × L did not yield improvements in validation performance for the two-dimensional latent space. All models were implemented in TensorFlow[24] and trained either on local workstations or using NVIDIA A100 GPUs on a high-performance computing cluster.

5. **Landscape Generation**

For each trained VAE model, we constructed a Direct Coupling Analysis (DCA)[16] model using the same input sequences used to train the VAE. The DCA model defines the probability of a sequence S of length L based on observed statistics of amino acids at individual positions A_i and pairs of positions (A_i, A_j). This probability is given by:

$$P(S) = \frac{1}{Z} \exp\left(\sum_{i<j} e_{ij}(A_i, A_j) + \sum h_i(A_i)\right) \quad (7)$$

Here, e_{ij} represents pairwise coupling parameters between positions i and j, while h_i corresponds to a local field term that captures the frequency of amino acids at position i. These parameters characterize the Boltzmann-like distribution over sequences and can be inferred using various methods[25–27].

To generate the landscape, we uniformly sampled coordinates (z_0, z_1) across the 2D latent space and passed them through the VAE decoder. The decoder produced a Softmax probability distribution over the 3Di states (Equation 6). From this, we extracted the maximum-probability sequence at each position as:

$$H(S^*) = -\sum_{1 \le i \le j \le L} e_{ij}(a_i, a_j) - \sum_{i=1}^{L} h_i(a_i) \quad (8)$$

where $S^* = a_i \ldots\ldots L$ and $a_i = arg \max_{a \in A} \; p(a|z)_i$

Each decoded sequence was then scored using the Hamiltonian derived from the DCA model. This Hamiltonian reflects the sequence's likelihood under the inferred coevolutionary constraints, enabling us to map evolutionary or structural plausibility across the latent space.

6. **Latent space entropy calculation**

To assess the uncertainty of sequence generation across the latent space, we computed the entropy landscape using the decoder's output distribution. For each coordinate in the latent space grid, the VAE decoder produces a Softmax distribution X over all amino acid symbols at every residue position. The average entropy per amino acid position at each coordinate provides a measure of variability or uncertainty in the generated sequences.

$$\widehat{H} = -\frac{1}{L} \sum_{i=1}^{L} \sum_{q=1}^{20} P\left(x_i^q\right) \log\left(P\left(x_i^q\right)\right) \quad (11)$$

7. **AlphaFold2 Structure Prediction and Performance Evaluation**

To evaluate the structural relevance of sequences generated by the VAE, we used Colabfold[5,28] to predict the 3D structures of both native and decoded sequences. Representative sequences were selected from across the latent space grid and from regions of interest identified in the clustering and contact prediction analyses. All predictions were performed with no structural templates and five recycles, ensuring that structure predictions were based solely on sequence information.

The AlphaFold-generated structures were used in three core evaluations. First, to assess how well the latent space captures true structural similarity, we computed pairwise TM-scores between structures of clustered sequences. Second, to validate contact map predictions from DCA, AlphaFold structures served as reference ground truth. Residue–residue contacts were defined based on a Cβ–Cβ distance cutoff of 8 Å (Cα for glycine) and were further validated using dynamic contacts from MD simulations. Finally, we examined whether 3Di sequences with favorable DCA Hamiltonian scores also produced well-folded structures.

8. Molecular Dynamics Simulations

The 3D structures predicted by AlphaFold2 were used as initial conformations for molecular dynamics (MD) simulations, which were performed using OpenMM[29]. Proteins were modeled using the AMBER14-all force field, and water molecules were represented using the TIP3P-FB model to construct fully solvated systems[30,31].

Each system was first subjected to energy minimization, followed by equilibration in the NVT ensemble (constant number of particles, volume, and temperature). This was then followed by a production run in the NPT ensemble (constant pressure and temperature) for a total duration of 1.2 microseconds at 298 K. To maintain thermodynamic stability, a Langevin Middle integrator was used to regulate the system temperature at 298 K, while pressure was controlled using OpenMM's Monte Carlo barostat to maintain 1 atm pressure[32].

9. Generation of Functional Sequences

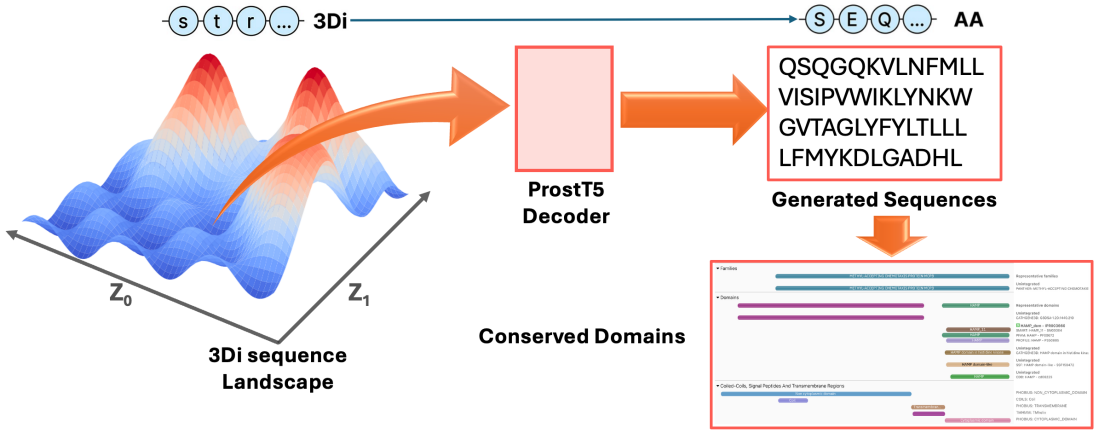

Fig 1: Schematic depiction of sampling from landscape encoding 3Di sequence, translating into real sequences via ProstT5 decoder and using the sequences for scoring and detecting structurally conserved domains.

To generate protein sequences with specific functional characteristics, we utilized the trained variational autoencoder (VAE) to decode latent coordinates into 3Di sequences (Fig 1). Latent coordinates were selected from regions of the landscape enriched for target domain annotations, such as the methyl-accepting chemotaxis (MAC) domain, based on their proximity to training sequences with known functions.

The decoder produced 3Di token sequences at each sampled coordinate, which were subsequently translated into amino acid sequences using the ProstT5 decoder model. To prioritize structurally conserved sequences, we selected latent coordinates based on their DCA Hamiltonian values. Low-Hamiltonian regions were favored for generating sequences hypothesized to retain native-like structure and functional domain features. The generated amino acid sequences were annotated using InterProScan[33] to assess the presence of conserved functional domains. For structural validation, AlphaFold2 was used to predict the 3D structures of selected generated sequences, and model confidence was assessed using the per-residue pLDDT scores.

Results

Structural Landscapes via ProstT5-3Di Pipeline

a. Generation of Structure-Informed Landscapes

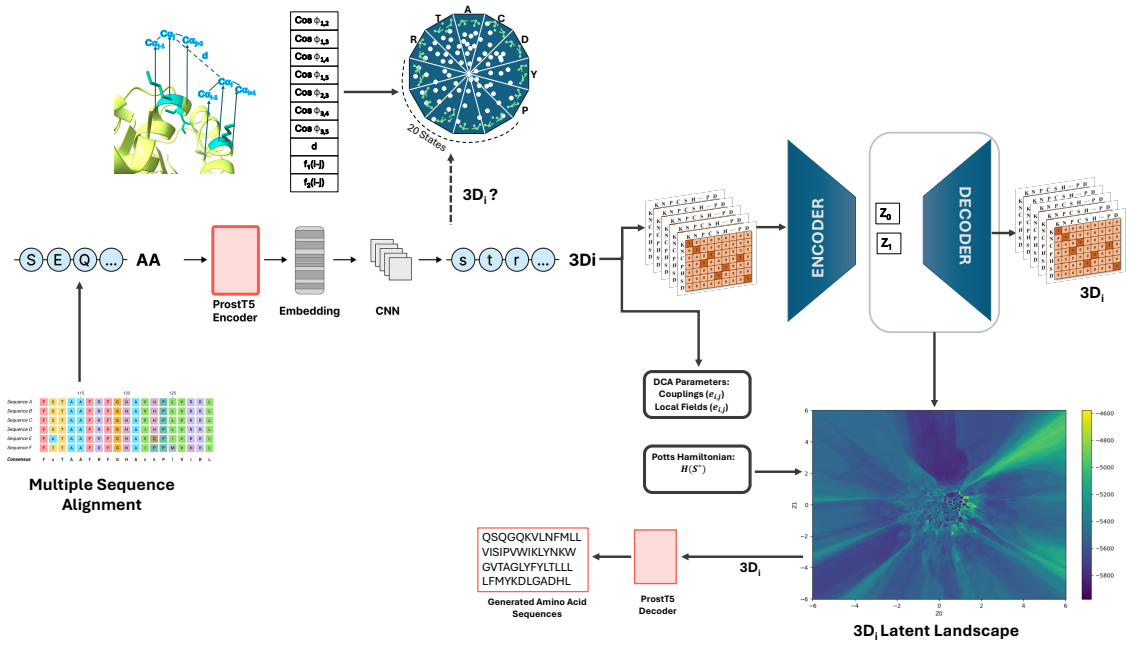

Fig 2: Overview of the ProstT5–3Di and VAE-based structural landscape generation pipeline. Multiple sequence alignments (MSAs) for various protein families are input into ProstT5, a pretrained language model that translates amino acid sequences into 3Di tokens representing local structural environments. MSA of these 3Di sequences are then processed through a variational autoencoder (VAE): the encoder maps sequences to latent coordinates (μ), while the decoder reconstructs them from this space. A grid of latent coordinates is sampled to generate maximum-likelihood sequences at each point. In parallel, Direct Coupling Analysis (DCA) is performed on the 3Di MSA to extract coevolutionary couplings and local fields, forming a Potts model Hamiltonian. This Hamiltonian score is used to assign an energy-based score to each decoded sequence, producing a 3D structural landscape.

We generated structure-informed sequence landscapes by integrating ProstT5, 3Di encoding, variational autoencoders (VAE), and direct coupling analysis (DCA) (Fig. 2). Multiple sequence alignments (MSAs) for each protein family were translated into 3Di tokens by ProstT5, encoding local residue geometry. The 3Di MSAs were processed by a VAE, which mapped sequences to latent coordinates and enabled sampling across the latent space. Sequences decoded at each point were scored using a DCA-derived Potts model Hamiltonian, producing an energy landscape where low-Hamiltonian regions correspond to structurally and functionally similar sequences.

b. Structural Clustering and Functional Annotation

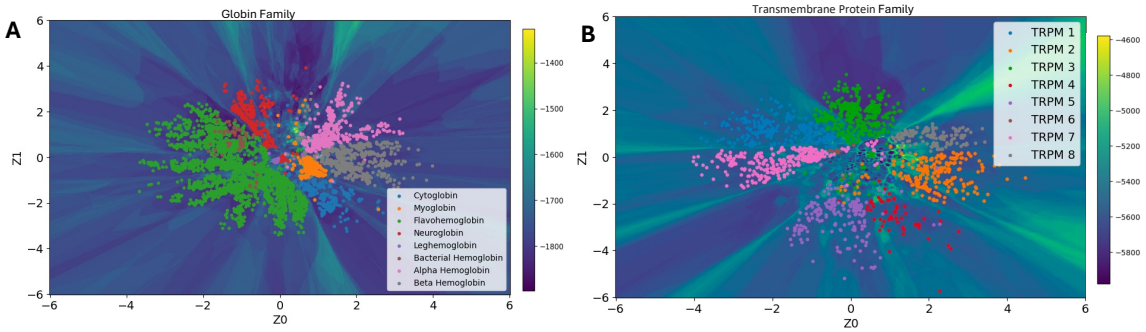

Fig 3: Functional and structural clustering of protein families in the 3Di-based latent landscape. (A) Globin family sequences cluster into distinct regions corresponding to UniProt-annotated functional types such as cytoglobin, myoglobin, and hemoglobin variants. (B) Transmembrane proteins (TRPM) cluster according to their known structural subdivisions—TRPM1/3, TRPM6/7, TRPM4/5, and TRPM2/8—indicating that the 3Di landscape captures architecture-driven relationships.

Across multiple training sets, we observed that the structural landscape effectively clusters 3Di sequences according to functional and structural similarities. For instance, in the Globin family (Fig. 3A), cluster boundaries align closely with UniProt functional annotations[34]. Similarly, kinase sequences—classified by structural and catalytic domains into nine groups (eight present in our dataset)—form distinct clusters in low-Hamiltonian basins (Supplementary Fig.)[35]. Moreover, local Hamiltonian barriers further subcluster the AGC kinase superfamily into subfamilies (Supplementary Fig.)[36]. These Hamiltonian-defined boundaries helps in classification based on structural distinctions.

A similar pattern emerges in Transmembrane Protein (TRPM) families, subdivided into four structurally similar groups—TRPM1/3, TRPM6/7, TRPM4/5, and TRPM2/8—which appear closely situated in the landscape (Fig. 3B)[37,38]. Thus, by focusing on structural states rather than amino acids alone, 3Di-based alignments yield landscapes where functionally related sequences naturally cluster.

c. Comparative Analysis of Structural and Sequence Landscapes

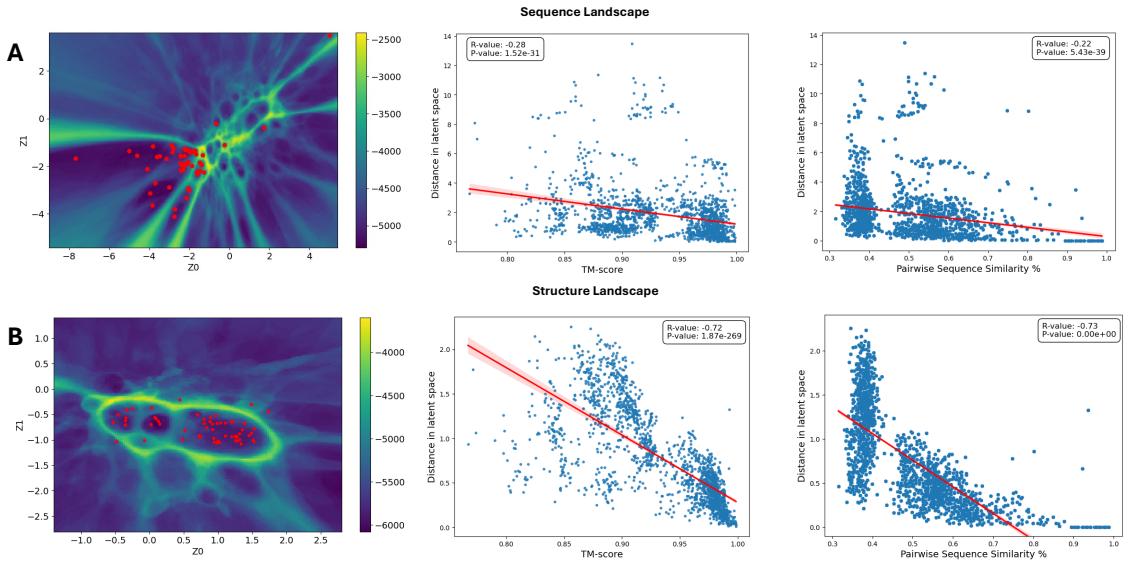

Fig 4: Correlation of latent space distances with sequence and structure similarity in the Malate Dehydrogenase (MDH) family. (A) In the amino acid–based latent space, red dots indicate sequences selected from within a Hamiltonian-defined cluster. While the landscape displays some structural grouping, the correlation between latent distance and TM-score is weak, and the association with sequence similarity is similarly low. (B) In contrast, the 3Di-based structural landscape shows much tighter clustering of structurally homologous sequences, with a strong negative correlation between latent space distance and TM-score.

To examine how structural similarity is captured these landscapes, we compared a standard amino acid (AA) MSA (Fig. 4A) with a 3Di MSA (Fig. 4B) for the Malate Dehydrogenase (MDH) family. Focusing on 50 sequences clustered within a Hamiltonian barrier boundary, we generated corresponding structures using AlphaFold[5]. Pairwise similarities were measured using Needleman–Wunsch (sequence) and TM-score (structure)[39,40]. A TM-score above 0.8 indicates structural homology, and we observed that the 3Di-based landscape exhibited tighter clustering of homologous sequences, showing a negative correlation of $r=-0.72$ between latent-space distance and TM-score[41]. Intriguingly, it also showed a higher correlation between sequence similarity and latent-space distance ($r=-0.73$) compared to the AA-based landscape ($r=-0.22$). This underscores the utility of structurally informed training, which can capture both sequence- and structure-level signals effectively[42,43].

d. Decoder–Sequence Relationship and Entropy Considerations

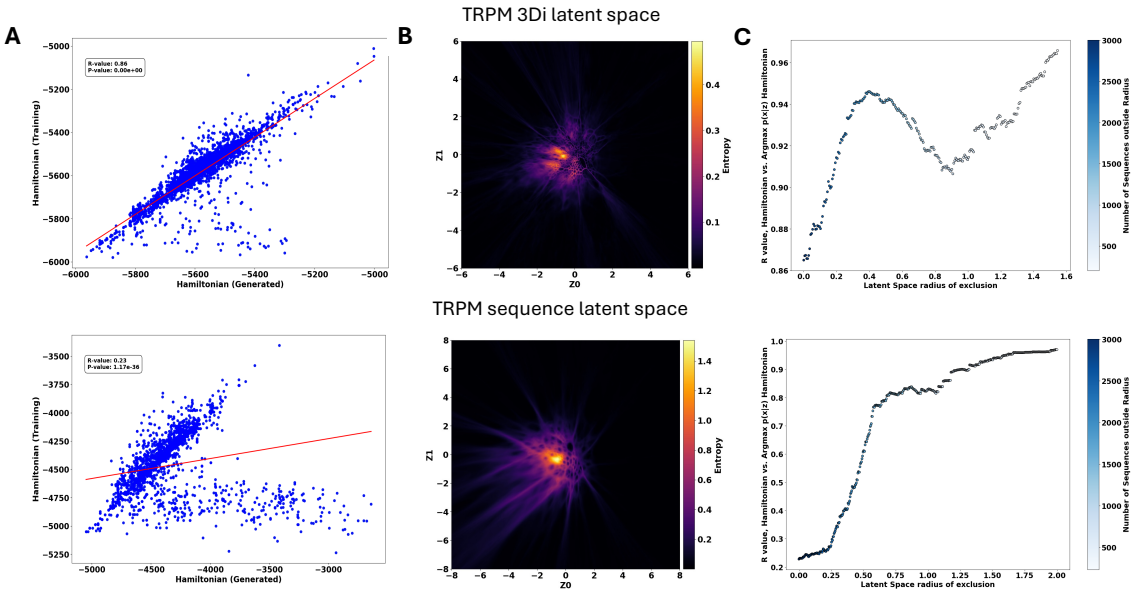

Fig 5: Decoder performance and entropy analysis across latent spaces for the TRPM family. (A) Comparison of DCA based Hamiltonian scores between input training sequences and sequences generated at the corresponding encoded μ-coordinate in both 3Di (top) and amino acid (bottom) landscapes. (B) Entropy of the decoder's output distribution at each latent coordinate, calculated using Equation (11). The 3Di landscape shows lower and more localized entropy compared to the broader, higher entropy in the amino acid latent space. (C) Correlation between input and generated Hamiltonians as a function of radial exclusion from the high-entropy latent space center. Excluding central, high-uncertainty sequences improves correlation in both landscapes, with the 3Di space achieving higher correlation even near the center. Color bar indicates the number of sequences remaining after each exclusion step.

To further investigate how encoded sequences relate to the decoder-produced landscapes, we embedded training sequences into both the sequence- and structure-based latent spaces and decoded the maximum-probability sequence at each encoded μ-coordinate. We then compared each decoded sequence's Hamiltonian to that of its original input. In the TRPM family (Fig. 5), both landscapes displayed a positive correlation, though the structural landscape exhibited a much stronger correlation (R=0.86) than the sequence landscape (R=0.23). This suggests that landscape generated by VAE over 3Di sequences can better capture distribution of input sequences as large variability, in structural sequences, of 3Di states will be present at structurally divergent regions, whereas amino acid sequences can vary at lesser structurally divergent regions too [20,44].

We also found that the correlation between input and generated Hamiltonians depends on the decoder's distributional variability, which peaks at the latent-space center (Fig. 5B). Entropy (Equation 8) is highest in these regions, reflecting greater uncertainty in the decoder's output. By excluding sequences located near these high-entropy coordinates, we consistently improved the correlation of LGL-generated (decoded) sequences (Fig. 5C). Crucially, the structural landscape retained lower entropy than its AA-based counterpart, reinforcing the notion that 3Di MSAs inject less noise and thereby smoothing the latent space of VAE. Although correlation levels varied by protein family (Supplementary Fig.), all tested families showed that moving outward from the high-entropy center markedly improved the correlation between input and generated Hamiltonians. Additionally, the structural landscape typically higher correlation in high-entropy regions; however, correlations in both landscapes converged to similar levels as the exclusion radius increased, indicating that sequences beyond the central "high-entropy" region are comparably well reconstructed in either landscape.

Evaluating Structural Contact Prediction with 3Di-DCA and Amino Acid MSAs

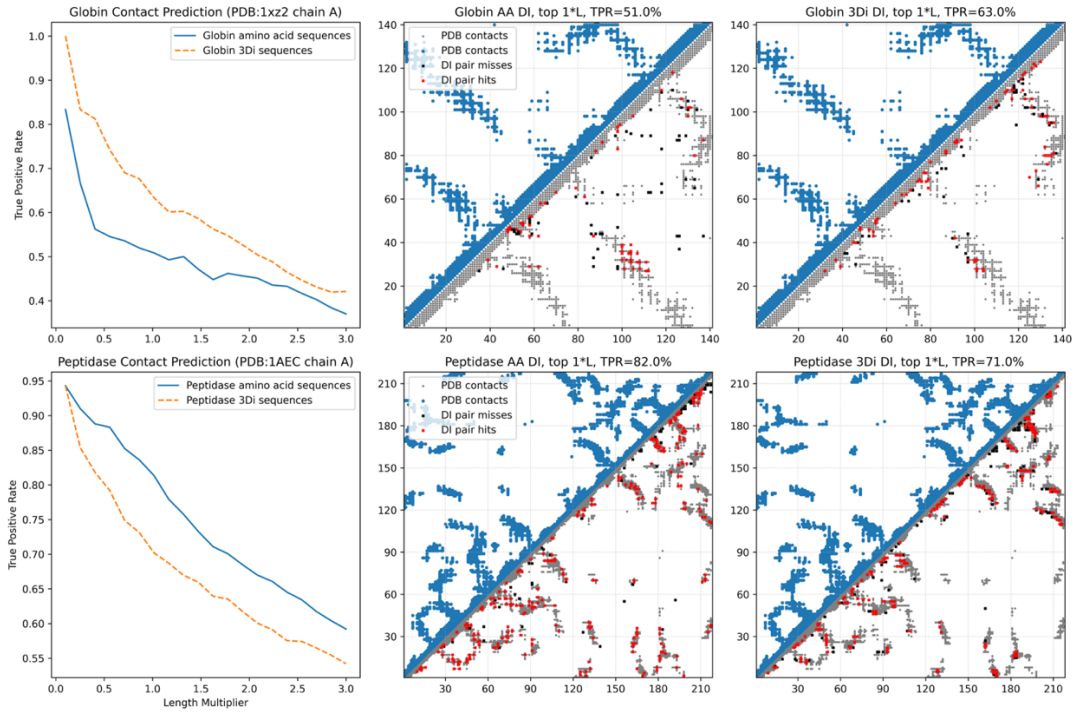

Fig 6. Comparison of contact predictions using 3Di and amino acid (AA) DCA respectively.
(left column) Precision of long-range contacts (|i–j| > 4) across top DI pairs (0.1L to 3L). (right columns) Top L direct interaction (DI) pairs for globin (top) and peptidase (bottom), long-range predictions (|i–j| > 4). Red: true positives; Black: false positives; gray/blue: PDB derived contacts (8 Å cutoff, all-atom). Globin PDB: 1xz2. Peptidase PDB: 1AEC.

To investigate the effectiveness of 3Di representations in residue-residue contact prediction, we conducted a comparative analysis of Direct Coupling Analysis (DCA) performed on 3Di-based and amino acid (AA) multiple sequence alignments (MSAs) for the globin and cysteine peptidase protein families. In the left column of Figure 6, we measure how well the DI pairs predict PDB derived residue contacts through a True Positive Rate over a range of the top ranked DI pairs. The top DI pairs have been well established to be predictive of residue-residue interaction in PDB structures, and in all examples, we see reasonable to excellent predictive power. In the case of globin, the 3Di sequence MSA clearly outperforms the original amino acid MSA, but this was not the case in the peptidase family. By looking at the contact maps themselves, we see that in the case of globin amino acid sequences there are substantially more false positives than in the 3Di sequences. While it is true that DI pairs can find conformational changes or allosteric effects[cite], we could not link the false positives to any documented result, and we may assume it to be noise.

The reversed result in the peptidase family may be because of the sample size, in that the peptidase MSA has 25,531 sequences available while the globin MSA only has 10,381. We reason that with a high number of diverse real sequences in the peptidase MSA, we have better estimates of the couplings which the 3Di may be damping and thus get more true positives. This also suggests that 3Di MSA might perform better with much less data, because each sequence contributes information to a fewer number of e_ij terms due to the reduced sequence

entropy (Figure 5), and this information is more directly relevant to known structures. To test this, we subsampled the globin and peptidase MSA and assessed contact prediction with smaller amounts of sequence data. We see that in general 3Di sequences provide better true positive rates compared to amino acid sequences especially at lower sequence counts, and this effect is even more pronounced in the globin family. We conclude that 3Di sequences might be beneficial for studying things like the effect of mutation on fitness or contact prediction in regimes of low data availability.

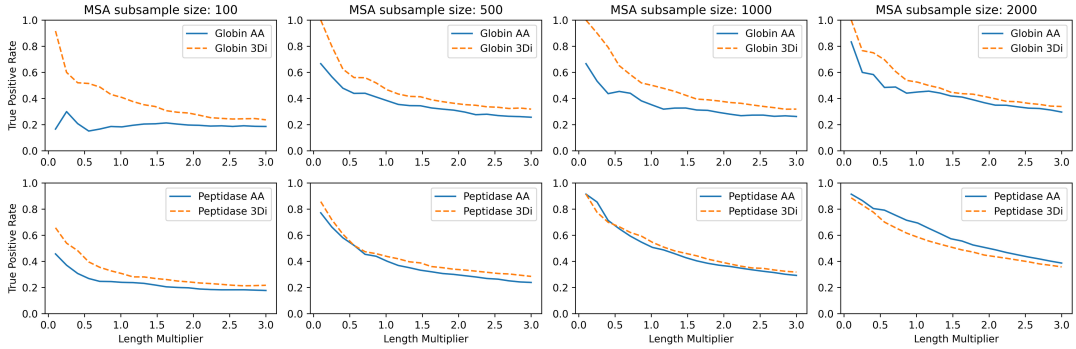

Fig 7. Comparison of contact predictions using 3Di and amino acid (AA) DCA across different MSA sizes. Precision of long-range contacts (|i–j| > 4) across top DI pairs (0.1L to 3L) against refrence structures: Globin PDB: 1xz2. Peptidase PDB: 1AEC. Each column represents an MSA subsampled from the corresponding full MSA, measured on AA and 3Di sequences.

3Di Landscapes Reveal Evolutionary Structure in Flaviviridae Glycoproteins

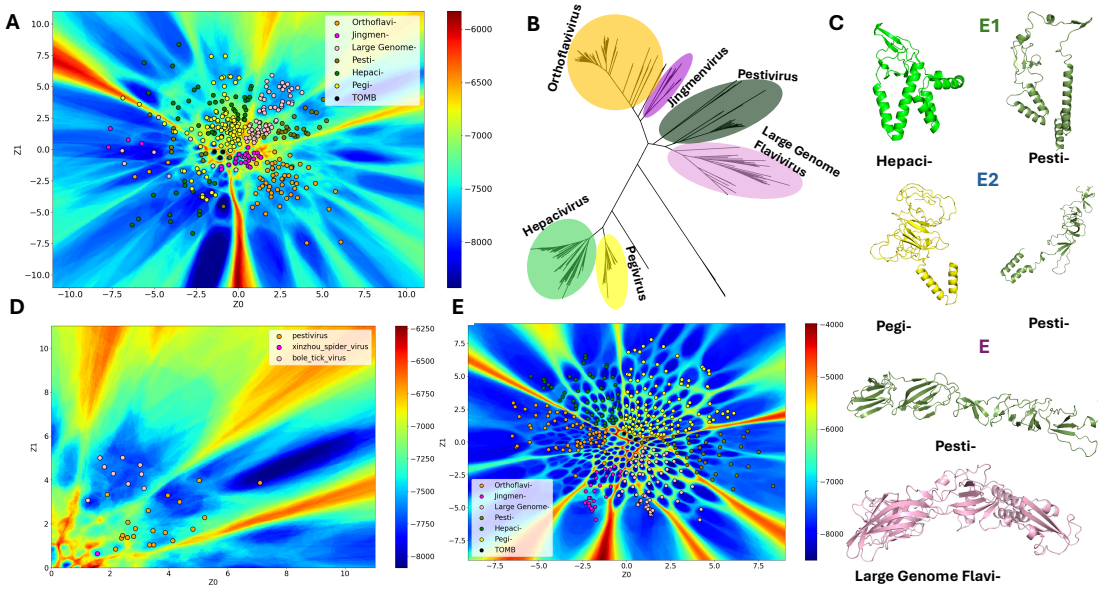

Fig 8. Landscape, phylogeny, and structural diversity of Flaviviridae glycoproteins.
(A) Latent space mapping of Flaviviridae 3Di sequences from the highly conserved NS5 region, colored by genera. (B) Phylogenetic reconstruction showing the placement of orthoflaviviruses, jingmenviruses, pestiviruses, LGF, hepaciviruses, and pegiviruses, consistent with latent space topology. (C) Representative predicted glycoprotein structures from each lineage, highlighting E1, E2, and E proteins across genera, including LGF-specific

architectures. (D) Latent space close-up of the LGF–pestivirus cluster, showing retention of similar structure features like RNase T2 in both the Bole Tick Virus group of LGF and pestiviruses, indicative of shared evolutionary history. (E) Latent space mapping of Flaviviridae amino acid sequences from the NS5 region.

The Flaviviridae[45] is a highly diverse family of enveloped positive-sense RNA viruses that includes major human pathogens such as dengue, Zika, and hepatitis C viruses, as well as animal pathogens like classical swine fever virus and bovine viral diarrhoea virus, and emerging threats such as West Nile, Alongshan, and Haseki tick viruses[46–48]. Currently classified into four genera—Orthoflavivirus, Pestivirus, Pegivirus, and Hepacivirus[49]—the family also encompasses recently discovered groups such as segmented jingmenviruses and large genome flaviviruses (LGFs), which challenge assumptions about RNA virus genome size and whose taxonomic positions remain unresolved. Traditional phylogenetics based on conserved proteins like RNA-dependent RNA polymerase provides limited insight into genome-wide relationships, as key determinants of virus biology[50,51]—such as glycoproteins essential for entry, host range, and immune evasion—are highly divergent and difficult to resolve with sequence-based[52,53] or classical structural approaches.

The latent space of amino acid sequences captures the phylogenetic relationships among sequences[18,43]; however, resolving relationships between remote homologs remains challenging. To address this, we converted the amino acid MSA of the highly conserved NS5 region of the Flaviviridae into their corresponding 3Di sequences. This transformation improves DCA signal resolution and enhances the capture of phylogenetic relationships among sequences.

Two distinct evolutionary lineages of Flaviviridae exist[54]. Lineage one has undergone extensive diversification. Genome segmentation occurred in the jingmenviruses, coinciding with divergence in their E glycoprotein[53], including the apparent loss of its fusion loop. Orthoflaviviruses acquired prM—a partner to the E glycoprotein—derived from a host chaperonin. The sister lineage gave rise to the LGF and Pestivirus subclades, consistent with the presence of jingmenviruses in the center of the landscape and their divergence into multiple subclades divided by Hamiltonian barriers(Fig. 7A). In this lineage, an ancestral virus acquired RNase T2, likely from a bacterial source, which has been retained in both the Bole Tick Virus group of the LGF and in pestiviruses(even in basal members like Xinzhou spider virus) (Fig. 7D).

In contrast, lineage two shows comparatively little diversity. The hepaciviruses and pegiviruses share similar genome organization, with high conservation in sequence and structure, as illustrated by their overlapping positions in the landscape (Fig. 7A). The presence of E1E2 is strictly associated with vertebrate viruses, suggesting that E1E2-dependent entry represents a molecular commitment to the vertebrate niche. This commitment appears to extend to species and even tissue tropism; for example, HCV efficiently infects only human hepatocytes, with E1E2 activity requiring a very specific set of entry factors. This specificity may underlie the limited divergence observed in the Hepacivirus/Pegivirus lineage.

Thus, the 3Di landscape provides a framework for resolving phylogenetic relationships among remote homologs capturing the contrasting evolutionary trajectories of the highly diversified orthoflavivirus–jingmenvirus–LGF–pestivirus lineage and the comparatively conserved hepacivirus–pegivirus lineage, thereby revealing structural and functional signatures of glycoprotein evolution.

3Di-based energy landscape barriers divide homologues based on dynamics similarity

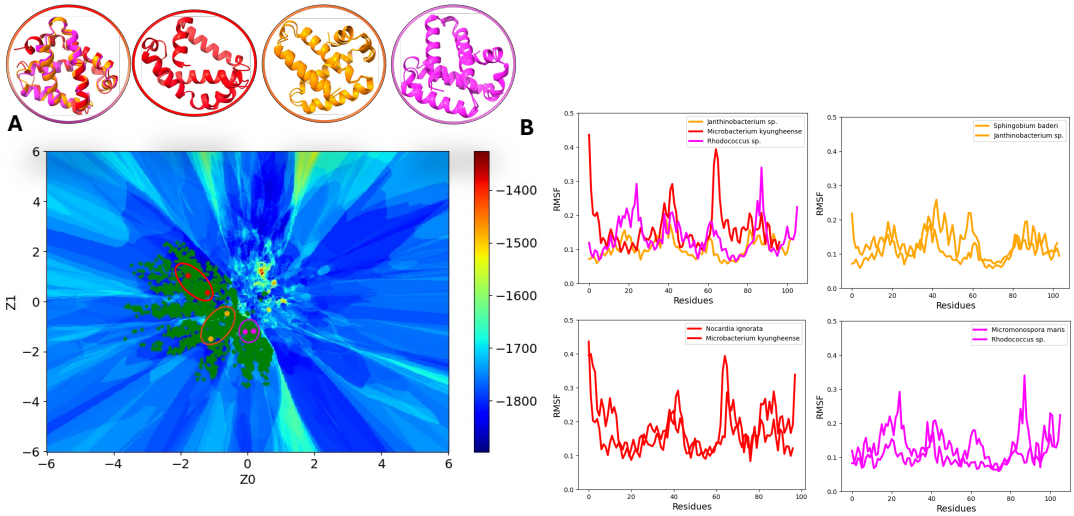

Fig 9: Dynamical similarity of flavohemoglobins reflected in 3Di-based latent landscape clustering. (A) 3Di latent landscape of globin family sequences colored by DCA based Hamiltonian score, with representative flavohemoglobins selected from three distinct clusters separated by Hamiltonian barriers (highlighted in red, orange, and purple). (B) Root mean square fluctuation (RMSF) profiles from 600 ns all-atom molecular dynamics simulations for AlphaFold-predicted structures of selected sequences. RMSF traces from proteins within the same cluster (e.g., red–purple, red–red) are nearly identical, indicating shared flexibility and dynamics. In contrast, proteins from separate clusters (e.g., orange–purple) show clear differences in RMSF, despite all being members of the same family.

Proteins that share similar folds and topological architecture often exhibit comparable dynamic behavior[55]. Since the 3Di-based landscape clusters sequences according to structural similarity, we investigated whether this clustering also captures dynamical similarity across members of the same family. To explore this, we selected representative Flavohemoglobins from distinct 3Di clusters separated by DCA based Hamiltonian barriers in the latent space (Fig. 9A). For each representative, 3D structures were predicted using AlphaFold2 and subjected to 600 ns all-atom molecular dynamics simulations. The root mean square fluctuation (RMSF) profiles of residues were then compared between pairs of proteins. We observed that proteins grouped within the same 3Di-defined cluster exhibited highly similar RMSF profiles, indicating shared dynamic behavior (Fig. B). In contrast, sequences from different clusters, even though all belonged to the same globin family, showed distinguishable fluctuations.

Discussion

Our work presents a unified, structure-aware framework for analyzing and generating protein sequences by integrating 3Di geometric encoding, variational autoencoders, and direct coupling analysis. By leveraging structural information through 3Di sequences, we construct latent landscapes that effectively organize proteins according to shared structural and functional features. Across diverse families, including globins, kinases, malate dehydrogenases, and transmembrane proteins, these landscapes reveal clear clustering patterns that align with known functional annotations, even for remote homologs with low sequence identity.

A key advance of our approach lies in coupling structure-informed latent spaces with coevolutionary constraints. The 3Di-based landscapes improve clustering and enhance contact prediction accuracy, outperforming traditional amino acid–based DCA, particularly in families where structural conservation outpaces sequence conservation. Our analysis of decoder entropy further highlights that 3Di landscapes reduce latent space noise, supporting more reliable sequence reconstruction and generative design. Importantly, we demonstrate that sequences generated from low-Hamiltonian regions of the structural landscape retain functional domain annotations, illustrating the potential of this framework for guided sequence design.

While our findings underscore the value of integrating structural representations into generative protein modeling, several limitations remain. The 3Di alphabet, while compact and informative, abstracts local geometry and may miss longer-range structural dependencies that contribute to the function. Similarly, the latent landscapes, though effective for clustering and design, depend on the quality of input MSAs and structural predictions. Incorporating explicit functional labels, experimental binding data, or dynamics information could further enhance the model's discriminative and generative power.

This framework opens avenues for exploring protein fitness landscapes, evolutionary trajectories, and the rational design of novel proteins with desired structural and functional properties. Extensions could include conditioning generative models on specific functional features, integrating physics-based scoring functions, or combining them with molecular dynamics to assess the stability and dynamics of generated sequences. Our work provides a foundation for structure-aware, data-driven protein modeling that bridges sequence, structure, and function in a unified generative paradigm.

References

(1) Barrio-Hernandez, I.; Yeo, J.; Jänes, J.; Mirdita, M.; Gilchrist, C. L. M.; Wein, T.; Varadi, M.; Velankar, S.; Beltrao, P.; Steinegger, M. Clustering Predicted Structures at the Scale of the Known Protein Universe. *Nature* **2023**, *622* (7983). https://doi.org/10.1038/s41586-023-06510-w.

(2) Chowdhury, R.; Bouatta, N.; Biswas, S.; Floristean, C.; Kharkar, A.; Roy, K.; Rochereau, C.; Ahdritz, G.; Zhang, J.; Church, G. M.; Sorger, P. K.; AlQuraishi, M. Single-Sequence Protein Structure Prediction Using a Language Model and Deep Learning. *Nat Biotechnol* **2022**, *40* (11). https://doi.org/10.1038/s41587-022-01432-w.

(3) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Dustin Schaeffer, R.; Millán, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A.; Pereira, J. H.; Rodrigues, A. V.; Van Dijk, A. A.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Christopher Garcia, K.; Grishin, N. V.; Adams, P. D.; Read, R. J.; Baker, D. Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network. *Science (1979)* **2021**, *373* (6557). https://doi.org/10.1126/science.abj8754.

(4) Varadi, M.; Anyango, S.; Deshpande, M.; Nair, S.; Natassia, C.; Yordanova, G.; Yuan, D.; Stroe, O.; Wood, G.; Laydon, A.; Zídek, A.; Green, T.; Tunyasuvunakool, K.; Petersen, S.; Jumper, J.; Clancy, E.; Green, R.; Vora, A.; Lutfi, M.; Figurnov, M.; Cowie, A.; Hobbs, N.; Kohli, P.; Kleywegt, G.; Birney, E.; Hassabis, D.; Velankar, S. AlphaFold Protein Structure Database: Massively Expanding the Structural Coverage of Protein-Sequence Space with High-Accuracy Models. *Nucleic Acids Res* **2022**, *50* (D1). https://doi.org/10.1093/nar/gkab1061.

(5) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873). https://doi.org/10.1038/s41586-021-03819-2.

(6) Wong, F.; Krishnan, A.; Zheng, E. J.; Stärk, H.; Manson, A. L.; Earl, A. M.; Jaakkola, T.; Collins, J. J. Benchmarking AlphaFold -enabled Molecular Docking Predictions for

Antibiotic Discovery . *Mol Syst Biol* **2022**, *18* (9). https://doi.org/10.15252/msb.202211081.

(7) Bordin, N.; Sillitoe, I.; Nallapareddy, V.; Rauer, C.; Lam, S. D.; Waman, V. P.; Sen, N.; Heinzinger, M.; Littmann, M.; Kim, S.; Velankar, S.; Steinegger, M.; Rost, B.; Orengo, C. AlphaFold2 Reveals Commonalities and Novelties in Protein Structure Space for 21 Model Organisms. *Commun Biol* **2023**, *6* (1). https://doi.org/10.1038/s42003-023-04488-9.

(8) Sommer, M. J.; Cha, S.; Varabyou, A.; Rincon, N.; Park, S.; Minkin, I.; Pertea, M.; Steinegger, M.; Salzberg, S. L. Structure-Guided Isoform Identification for the Human Transcriptome. *Elife* **2022**, *11*. https://doi.org/10.7554/ELIFE.82556.

(9) Paysan-Lafosse, T.; Blum, M.; Chuguransky, S.; Grego, T.; Pinto, B. L.; Salazar, G. A.; Bileschi, M. L.; Bork, P.; Bridge, A.; Colwell, L.; Gough, J.; Haft, D. H.; Letunić, I.; Marchler-Bauer, A.; Mi, H.; Natale, D. A.; Orengo, C. A.; Pandurangan, A. P.; Rivoire, C.; Sigrist, C. J. A.; Sillitoe, I.; Thanki, N.; Thomas, P. D.; Tosatto, S. C. E.; Wu, C. H.; Bateman, A. InterPro in 2022. *Nucleic Acids Res* **2023**, *51* (D1). https://doi.org/10.1093/nar/gkac993.

(10) Richardson, L.; Allen, B.; Baldi, G.; Beracochea, M.; Bileschi, M. L.; Burdett, T.; Burgin, J.; Caballero-Perez, J.; Cochrane, G.; Colwell, L. J.; Curtis, T.; Escobar-Zepeda, A.; Gurbich, T. A.; Kale, V.; Korobeynikov, A.; Raj, S.; Rogers, A. B.; Sakharova, E.; Sanchez, S.; Wilkinson, D. J.; Finn, R. D. MGnify: The Microbiome Sequence Data Analysis Resource in 2023. *Nucleic Acids Res* **2023**, *51* (1 D). https://doi.org/10.1093/nar/gkac1080.

(11) Vanni, C.; Schechter, M. S.; Acinas, S. G.; Barberán, A.; Buttigieg, P. L.; Casamayor, E. O.; Delmont, T. O.; Duarte, C. M.; Eren, A. M.; Finn, R. D.; Kottmann, R.; Mitchell, A.; Sanchez, P.; Siren, K.; Steinegger, M.; Glöckner, F. O.; Fernandez-Guerra, A. Unifying the Known and Unknown Microbial Coding Sequence Space. *Elife* **2022**, *11*. https://doi.org/10.7554/eLife.67667.

(12) Lundin, D.; Poole, A. M.; Sjöberg, B. M.; Högbom, M. Use of Structural Phylogenetic Networks for Classification of the Ferritin-like Superfamily. *Journal of Biological Chemistry* **2012**, *287* (24). https://doi.org/10.1074/jbc.M112.367458.

(13) van Kempen, M.; Kim, S. S.; Tumescheit, C.; Mirdita, M.; Lee, J.; Gilchrist, C. L. M.; Söding, J.; Steinegger, M. Fast and Accurate Protein Structure Search with Foldseek. *Nat Biotechnol* **2024**, *42* (2). https://doi.org/10.1038/s41587-023-01773-0.

(14) Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; Bhowmik, D.; Rost, B. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Trans Pattern Anal Mach Intell* **2022**, *44* (10). https://doi.org/10.1109/TPAMI.2021.3095381.

(15) Heinzinger, M.; Weissenow, K.; Gomez Sanchez, J.; Henkel, A.; Mirdita, M.; Steinegger, M.; Rost, & B.; Heinzinger, M.; Weissenow, K.; Gomez Sanchez, J.; Henkel, A.; Mirdita, M.; Steinegger, M.; Prostt5, R. Bilingual Language Model for Protein Sequence and Structure. *bioRxiv* **2024**.

(16) Morcos, F.; Pagnani, A.; Lunt, B.; Bertolino, A.; Marks, D. S.; Sander, C.; Zecchina, R.; Onuchic, J. N.; Hwa, T.; Weigt, M. Direct-Coupling Analysis of Residue Coevolution Captures Native Contacts across Many Protein Families. *Proc Natl Acad Sci U S A* **2011**, *108* (49). https://doi.org/10.1073/pnas.1111471108.

(17) Kingma, D. P.; Welling, M. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*; 2014. https://doi.org/10.61603/ceas.v2i1.33.

(18) Ziegler, C.; Martin, J.; Sinner, C.; Morcos, F. Latent Generative Landscapes as Maps of Functional Diversity in Protein Sequence Space. *Nat Commun* **2023**, *14* (1). https://doi.org/10.1038/s41467-023-37958-z.

(19) Anton, B.; Besalú, M.; Fornes, O.; Bonet, J.; Molina, A.; Molina-Fernandez, R.; De Las Cuevas, G.; Fernandez-Fuentes, N.; Oliva, B. On the Use of Direct-Coupling Analysis with a Reduced Alphabet of Amino Acids Combined with Super-Secondary Structure Motifs for Protein Fold Prediction. *NAR Genom Bioinform* **2021**, *3* (2). https://doi.org/10.1093/nargab/lqab027.

(20) Lesk, A. M.; Chothia, C. How Different Amino Acid Sequences Determine Similar Protein Structures: The Structure and Evolutionary Dynamics of the Globins. *J Mol Biol* **1980**, *136* (3). https://doi.org/10.1016/0022-2836(80)90373-3.

(21) Finn, R. D.; Clements, J.; Eddy, S. R. HMMER Web Server: Interactive Sequence Similarity Searching. *Nucleic Acids Res* **2011**, *39* (SUPPL. 2). https://doi.org/10.1093/nar/gkr367.

(22) Kalaivani, R.; Reema, R.; Srinivasan, N. Recognition of Sites of Functional Specialisation in All Known Eukaryotic Protein Kinase Families. *PLoS Comput Biol* **2018**, *14* (2). https://doi.org/10.1371/journal.pcbi.1005975.

(23) Mistry, J.; Chuguransky, S.; Williams, L.; Qureshi, M.; Salazar, G. A.; Sonnhammer, E. L. L.; Tosatto, S. C. E.; Paladin, L.; Raj, S.; Richardson, L. J.; Finn, R. D.; Bateman, A.

Pfam: The Protein Families Database in 2021. *Nucleic Acids Res* **2021**, *49* (D1). https://doi.org/10.1093/nar/gkaa913.

(24) Rucci, M.; Casile, A. *{TensorFlow}: Large-Scale Machine Learning on Heterogeneous Systems}*. Network: Computation in Neural Systems.

(25) Trinquier, J.; Uguzzoni, G.; Pagnani, A.; Zamponi, F.; Weigt, M. Efficient Generative Modeling of Protein Sequences Using Simple Autoregressive Models. *Nat Commun* **2021**, *12* (1). https://doi.org/10.1038/s41467-021-25756-4.

(26) Figliuzzi, M.; Barrat-Charlaix, P.; Weigt, M. How Pairwise Coevolutionary Models Capture the Collective Residue Variability in Proteins? *Mol Biol Evol* **2018**, *35* (4). https://doi.org/10.1093/molbev/msy007.

(27) Ekeberg, M.; Lövkvist, C.; Lan, Y.; Weigt, M.; Aurell, E. Improved Contact Prediction in Proteins: Using Pseudolikelihoods to Infer Potts Models. *Phys Rev E Stat Nonlin Soft Matter Phys* **2013**, *87* (1). https://doi.org/10.1103/PhysRevE.87.012707.

(28) Mirdita, M.; Schütze, K.; Moriwaki, Y.; Heo, L.; Ovchinnikov, S.; Steinegger, M. ColabFold: Making Protein Folding Accessible to All. *Nat Methods* **2022**, *19* (6). https://doi.org/10.1038/s41592-022-01488-1.

(29) Eastman, P.; Swails, J.; Chodera, J. D.; McGibbon, R. T.; Zhao, Y.; Beauchamp, K. A.; Wang, L. P.; Simmonett, A. C.; Harrigan, M. P.; Stern, C. D.; Wiewiora, R. P.; Brooks, B. R.; Pande, V. S. OpenMM 7: Rapid Development of High Performance Algorithms for Molecular Dynamics. *PLoS Comput Biol* **2017**, *13* (7). https://doi.org/10.1371/journal.pcbi.1005659.

(30) Wang, L. P.; Martinez, T. J.; Pande, V. S. Building Force Fields: An Automatic, Systematic, and Reproducible Approach. *Journal of Physical Chemistry Letters* **2014**, *5* (11). https://doi.org/10.1021/jz500737m.

(31) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J Chem Phys* **1983**, *79* (2). https://doi.org/10.1063/1.445869.

(32) Martyna, G. J.; Tobias, D. J.; Klein, M. L. Constant Pressure Molecular Dynamics Algorithms. *J Chem Phys* **1994**, *101* (5). https://doi.org/10.1063/1.467468.

(33) Quevillon, E.; Silventoinen, V.; Pillai, S.; Harte, N.; Mulder, N.; Apweiler, R.; Lopez, R. InterProScan: Protein Domains Identifier. *Nucleic Acids Res* **2005**, *33* (SUPPL. 2). https://doi.org/10.1093/nar/gki442.

(34) Bateman, A.; Martin, M. J.; Orchard, S.; Magrane, M.; Ahmad, S.; Alpi, E.; Bowler-Barnett, E. H.; Britto, R.; Bye-A-Jee, H.; Cukura, A.; Denny, P.; Dogan, T.; Ebenezer, T. G.; Fan, J.; Garmiri, P.; da Costa Gonzales, L. J.; Hatton-Ellis, E.; Hussein, A.; Ignatchenko, A.; Insana, G.; Ishtiaq, R.; Joshi, V.; Jyothi, D.; Kandasaamy, S.; Lock, A.; Luciani, A.; Lugaric, M.; Luo, J.; Lussi, Y.; MacDougall, A.; Madeira, F.; Mahmoudy, M.; Mishra, A.; Moulang, K.; Nightingale, A.; Pundir, S.; Qi, G.; Raj, S.; Raposo, P.; Rice, D. L.; Saidi, R.; Santos, R.; Speretta, E.; Stephenson, J.; Totoo, P.; Turner, E.; Tyagi, N.; Vasudev, P.; Warner, K.; Watkins, X.; Zaru, R.; Zellner, H.; Bridge, A. J.; Aimo, L.; Argoud-Puy, G.; Auchincloss, A. H.; Axelsen, K. B.; Bansal, P.; Baratin, D.; Batista Neto, T. M.; Blatter, M. C.; Bolleman, J. T.; Boutet, E.; Breuza, L.; Gil, B. C.; Casals-Casas, C.; Echioukh, K. C.; Coudert, E.; Cuche, B.; de Castro, E.; Estreicher, A.; Famiglietti, M. L.; Feuermann, M.; Gasteiger, E.; Gaudet, P.; Gehant, S.; Gerritsen, V.; Gos, A.; Gruaz, N.; Hulo, C.; Hyka-Nouspikel, N.; Jungo, F.; Kerhornou, A.; Le Mercier, P.; Lieberherr, D.; Masson, P.; Morgat, A.; Muthukrishnan, V.; Paesano, S.; Pedruzzi, I.; Pilbout, S.; Pourcel, L.; Poux, S.; Pozzato, M.; Pruess, M.; Redaschi, N.; Rivoire, C.; Sigrist, C. J. A.; Sonesson, K.; Sundaram, S.; Wu, C. H.; Arighi, C. N.; Arminski, L.; Chen, C.; Chen, Y.; Huang, H.; Laiho, K.; McGarvey, P.; Natale, D. A.; Ross, K.; Vinayaka, C. R.; Wang, Q.; Wang, Y.; Zhang, J. UniProt: The Universal Protein Knowledgebase in 2023. *Nucleic Acids Res* **2023**, *51* (D1). https://doi.org/10.1093/nar/gkac1052.

(35) Bradley, D.; Beltrao, P. Evolution of Protein Kinase Substrate Recognition at the Active Site. *PLoS Biol* **2019**, *17* (6). https://doi.org/10.1371/journal.pbio.3000341.

(36) Pearce, L. R.; Komander, D.; Alessi, D. R. The Nuts and Bolts of AGC Protein Kinases. *Nature Reviews Molecular Cell Biology*. 2010. https://doi.org/10.1038/nrm2822.

(37) Ciaglia, T.; Vestuto, V.; Bertamino, A.; González-Muñiz, R.; Gómez-Monterrey, I. On the Modulation of TRPM Channels: Current Perspectives and Anticancer Therapeutic Implications. *Frontiers in Oncology*. 2023. https://doi.org/10.3389/fonc.2022.1065935.

(38) Nilius, B.; Owsianik, G. The Transient Receptor Potential Family of Ion Channels. *Genome Biology*. 2011. https://doi.org/10.1186/gb-2011-12-3-218.

(39) Zhang, Y.; Skolnick, J. TM-Align: A Protein Structure Alignment Algorithm Based on the TM-Score. *Nucleic Acids Res* **2005**, *33* (7). https://doi.org/10.1093/nar/gki524.

(40) Needleman, S. B.; Wunsch, C. D. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *J Mol Biol* **1970**, *48* (3). https://doi.org/10.1016/0022-2836(70)90057-4.

(41) Xu, J.; Zhang, Y. How Significant Is a Protein Structure Similarity with TM-Score = 0.5? *Bioinformatics* **2010**, *26* (7). https://doi.org/10.1093/bioinformatics/btq066.

(42) Tian, H.; Jiang, X.; Xiao, S.; La Force, H.; Larson, E. C.; Tao, P. LAST: Latent Space-Assisted Adaptive Sampling for Protein Trajectories. *J Chem Inf Model* **2023**, *63* (1). https://doi.org/10.1021/acs.jcim.2c01213.

(43) Ding, X.; Zou, Z.; Brooks, C. L. Deciphering Protein Evolution and Fitness Landscapes with Latent Space Models. *Nat Commun* **2019**, *10* (1). https://doi.org/10.1038/s41467-019-13633-0.

(44) Puente-Lelievre, C.; Malik, A. J.; Douglas, J.; Ascher, D.; Baker, M.; Allison, J.; Poole, A.; Lundin, D.; Fullmer, M.; Bouckert, R.; Kim, H.; Steinegger, M.; Matzke, N. Tertiary-Interaction Characters Enable Fast, Model-Based Structural Phylogenetics beyond the Twilight Zone. *bioRxiv* **2024**.

(45) Simmonds, P.; Becher, P.; Bukh, J.; Gould, E. A.; Meyers, G.; Monath, T.; Muerhoff, S.; Pletnev, A.; Rico-Hesse, R.; Smith, D. B.; Stapleton, J. T. ICTV Virus Taxonomy Profile: Flaviviridae. *Journal of General Virology* **2017**, *98* (1). https://doi.org/10.1099/jgv.0.000672.

(46) Kartashov, M. Y.; Gladysheva, A. V.; Shvalov, A. N.; Tupota, N. L.; Chernikova, A. A.; Ternovoi, V. A.; Loktev, V. B. Novel Flavi-like Virus in Ixodid Ticks and Patients in Russia. *Ticks Tick Borne Dis* **2023**, *14* (2). https://doi.org/10.1016/j.ttbdis.2022.102101.

(47) Wang, Z.-D.; Wang, B.; Wei, F.; Han, S.-Z.; Zhang, L.; Yang, Z.-T.; Yan, Y.; Lv, X.-L.; Li, L.; Wang, S.-C.; Song, M.-X.; Zhang, H.-J.; Huang, S.-J.; Chen, J.; Huang, F.-Q.; Li, S.; Liu, H.-H.; Hong, J.; Jin, Y.-L.; Wang, W.; Zhou, J.-Y.; Liu, Q. A New Segmented Virus Associated with Human Febrile Illness in China. *New England Journal of Medicine* **2019**, *380* (22). https://doi.org/10.1056/nejmoa1805068.

(48) Hubálek, Z.; Halouzka, J. West Nile Fever - A Reemerging Mosquito-Borne Viral Disease in Europe. *Emerg Infect Dis* **1999**, *5* (5). https://doi.org/10.3201/eid0505.990505.

(49) Postler, T. S.; Beer, M.; Blitvich, B. J.; Bukh, J.; de Lamballerie, X.; Drexler, J. F.; Imrie, A.; Kapoor, A.; Karganova, G. G.; Lemey, P.; Lohmann, V.; Simmonds, P.; Smith, D. B.; Stapleton, J. T.; Kuhn, J. H. Renaming of the Genus Flavivirus to Orthoflavivirus and Extension of Binomial Species Names within the Family Flaviviridae. *Archives of Virology*. 2023. https://doi.org/10.1007/s00705-023-05835-1.

(50) Shi, M.; Lin, X.-D.; Vasilakis, N.; Tian, J.-H.; Li, C.-X.; Chen, L.-J.; Eastwood, G.; Diao, X.-N.; Chen, M.-H.; Chen, X.; Qin, X.-C.; Widen, S. G.; Wood, T. G.; Tesh, R. B.; Xu, J.; Holmes, E. C.; Zhang, Y.-Z. Divergent Viruses Discovered in Arthropods and Vertebrates Revise the Evolutionary History of the Flaviviridae and Related Viruses. *J Virol* **2016**, *90* (2). https://doi.org/10.1128/jvi.02036-15.

(51) Paraskevopoulou, S.; Käfer, S.; Zirkel, F.; Donath, A.; Petersen, M.; Liu, S.; Zhou, X.; Drosten, C.; Misof, B.; Junglen, S. Viromics of Extant Insect Orders Unveil the Evolution of the Flavi-like Superfamily. *Virus Evol* **2021**, *7* (1). https://doi.org/10.1093/ve/veab030.

(52) Petrone, M. E.; Grove, J.; Mélade, J.; Mifsud, J. C. O.; Parry, R. H.; Marzinelli, E. M.; Holmes, E. C. A ~40-Kb Flavi-like Virus Does Not Encode a Known Error-Correcting Mechanism. *Proceedings of the National Academy of Sciences* **2024**, *121* (30). https://doi.org/10.1073/pnas.2403805121.

(53) Garry, C. E.; Garry, R. F. Proteomics Computational Analyses Suggest That the Envelope Glycoproteins of Segmented Jingmen Flavi-like Viruses Are Class II Viral Fusion Proteins (β-Penetrenes) with Mucin-like Domains. *Viruses* **2020**, *12* (3). https://doi.org/10.3390/v12030260.

(54) Mifsud, J. C. O.; Lytras, S.; Oliver, M. R.; Toon, K.; Costa, V. A.; Holmes, E. C.; Grove, J. Mapping Glycoprotein Structure Reveals Flaviviridae Evolutionary History. *Nature* **2024**, *633* (8030), 695–703. https://doi.org/10.1038/s41586-024-07899-8.

(55) Keskin, O.; Jernigan, R. L.; Bahar, I. Proteins with Similar Architecture Exhibit Similar Large-Scale Dynamic Behavior. *Biophys J* **2000**, *78* (4). https://doi.org/10.1016/S0006-3495(00)76756-7.