Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

From single-sequence structure prediction to protein fitness landscape through a composable, epistasis-aware mutation atlas

Yunxin Xu

xyx@zuel.edu.cn

Zhongnan University of Economics and Law https://orcid.org/0000-0002-9981-2097

Weizhe Wang

Tsinghua University

Zimu Yu

Zhongnan University of Economics and Law

Endi Yang

Zhongnan University of Economics and Law

Ziyu Shi

Zhongnan University of Economics and Law

Shize Yu

Huazhong University of Science and Technology https://orcid.org/0009-0005-0448-9055

Jian Hu

Tsinghua University

Haipeng Gong

Tsinghua University https://orcid.org/0000-0002-5532-1640

Article

Keywords:

Posted Date: September 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11164331/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Yes there is potential Competing Interest. Yunxin Xu and Zimu Yu have submitted a patent application based on this Epistasis work. The remaining authors declare no

competing interests.

From single-sequence structure prediction to protein fitness landscape through a composable, epistasis-aware mutation atlas

Weizhe Wang [1,2,3,4,†], Zimu Yu [5,†], Endi Yang [5], Ziyu Shi [5], Shize Yu [6], Jian Hu [1,2,3,4], Yunxin Xu [5,*], and Haipeng Gong [1,2,3,4,*]

[1]MOE Key Laboratory of Bioinformatics, School of Life Sciences, Tsinghua University, Beijing 100084, China
[2]State Key Laboratory of Green Manufacturing, Tsinghua University, Beijing 100084, China
[3]Beijing Frontier Research Center for Biological Structure, Tsinghua University, Beijing 100084, China
[4]Center of Life Sciences and Artificial Intelligence, School of Life Sciences, Tsinghua University, Beijing 100084, China
[5]School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan 430073, China
[6]School of Software Engineering, Huazhong University of Science and Technology, Wuhan 430074, China
[†]These authors contributed equally.
[*]Corresponding author. E-mail: hgong@tsinghua.edu.cn or xyx@zuel.edu.cn

Abstract

Mapping the multi-mutant fitness landscape is vital to protein engineering, but is challenging due to the vast combinatorial sequence space awaiting exploration. A central difficulty lies in the accurate and efficient modeling of non-additive epistatic effects among individual mutations, which partially arise from the physical inter-residue interactions prescribed by the protein structure. Existing fitness predictors usually perform well on single mutants but become less powerful for higher-order mutants, due to the lack of explicitly considering the relationship between sequence, structure and function of the target protein. Here, we present an end-to-end framework named Cerebra-Epistasis, which couples a single-sequence structure predictor that explicitly endows the structure awareness beyond the conventional sequence-fitness mapping with a downstream fitness prediction network that deliberately models the non-linear epistatic effects beyond the traditional additive terms. When evaluated across diverse assays, Cerebra-Epistasis outperforms the other state-of-the-art baselines in multi-mutant fitness prediction, with enhanced advantage over increasing mutation orders. Moreover, our special design on the epistasis modeling allows reliable extrapolation from low-order mutant data to unseen higher-order combinations, enabling one-shot inference of the overall mutation atlas from the starting sequence, a benefit that supposedly introduces three orders of magnitude acceleration in the landscape-scale prediction.

Introduction

Engineering protein function by sequence variation has broad biotechnological impacts, and has been extensively applied in practical fields including enzyme engineering and therapeutic development. Experimental methods like deep mutational scanning (DMS) can only capture a tiny fraction of the full sequence combinations from a wild-type protein [1, 2]. This deficit leaves the large-scale unmeasured mutational spectrum to be explored and predicted by computational methods. To guide protein optimization, a computational model is supposed to accurately map the multi-mutant fitness landscape, particularly the epistatic effects between individual mutations, while supporting efficient inference at the landscape scale.

Early evolutionary models infer mutational effects from the multiple sequence alignment (MSA) using sitewise conservation or pairwise sequence covariation [3]. Subsequently developed protein language models (PLMs) enable zero-shot mutation scoring from representations learned on large-scale sequence data [4, 5]. In these methods, epistatic effects are either completely neglected by the additive processing of sitewise scores or oversimplified as sparse pairwise terms using graphical models. Sparse regression methods, such as LASSO, extend such pairwise decomposition by directly fitting selected low-order interaction terms to measured fitness landscapes [6], but remain powerless in inferring unseen mutants due to the limited model-specified terms. More recent supervised methods begin to learn the nonlinear projection from complete variant sequences to measured fitness. ProteinNPT combines protein representations with available property measurements via a non-parametric transformer, and thus supports prediction in label-scarce and multi-property settings [7]. By combining PLM embeddings with a regression model, EVOLVEpro iteratively updates the model using small batches of experimental measurements to guide protein optimization [8]. MULTI-evolve focuses more directly on higher-order mutant design, by training a nonlinear predictor on strategically measured single- and double-mutant variants and then applying it to higher-order combinations [9].

Despite versatile objectives, these methods directly predict the total fitness, and more importantly, require each candidate variant to be represented and evaluated separately, which leads to the scaling of storage, GPU memory and computation time with the candidate size during the construction of a large-scale fitness landscape. One possible solution is to construct separate reusable representations of single-mutation effects and epistatic effects, respectively, which not only allows instant derivation of the total fitness through direct assembly but also retains the flexibility of results across mutation orders. The epistatic representation should be able to reflect the change of pairwise effects upon the alteration of mutational background, so as to capture the prominent three-way and higher-order epistasis revealed in previous research [6, 10]. Thus, the remaining challenge is to combine accurate fitness prediction, explicit epistasis modeling across varying mutation orders, and efficient landscape-scale inference within a unified framework.

The 3D structure of a protein provides physical context for epistatic interactions: residues that are distant in sequence can be proximal in space, giving rise to coupled effects. PLMs and MSA-based methods capture evolutionary constraints but do not explicitly represent structural information such as local packing, long-range contacts, and/or side-chain geometry. Structure-based inverse-folding scores offer complementary information to sequence-based models, and multimodal approaches that integrate sequence, MSA and structure have been shown to improve mutation-effect prediction across

diverse assay types [11, 12]. Building on these efforts, Kermut combines sequence and structural information through a composite Gaussian process kernel for supervised variant-effect prediction [13]. SPIRED-Fitness moves further by coupling a single-sequence structure predictor to a fitness model and training both components end to end [14]. Together, by demonstrating that structural context can inform mutation-effect prediction, these studies motivate the integration of structure awareness with explicit epistasis modeling. However, realizing such an integration scheme within an end-to-end fitness framework requires a fast, reliable structure predictor that supports end-to-end optimization at an acceptable computational cost.

MSA-based structure predictors such as AlphaFold 2 require extensive preprocessing and substantial computation [15]. Single-sequence models such as ESMFold and OmegaFold circumvent the MSA-search bottleneck to accelerate inference, but still require computation and memory intensive training [16, 17], which hinders end-to-end training with downstream fitness models. SPIRED-Fitness overcomes this problem by using a lightweight single-sequence predictor SPIRED, which represents structures mainly through C_α coordinates but has to rely on a separate pipeline for all-atom reconstruction [14]. An ideal structure-aware fitness framework therefore demands a structure predictor that combines high structural accuracy, single-sequence efficiency, side-chain-resolved representations, and sufficiently low computational cost for end-to-end training.

In this study, we present Cerebra-Epistasis, an end-to-end structure-aware framework for efficient construction of the protein fitness landscape. Its structure prediction module, Cerebra-Seq, is an accurate and efficient single-sequence model derived from the original MSA-based predictor Cerebra [18]. Cerebra-Seq produces side-chain-resolved structural representations and supports end-to-end training with the downstream fitness prediction network, a design that provides explicit 3D structure awareness for each target protein. The downstream fitness predictor decomposes multi-mutant fitness into additive single-mutation components and an independent epistatic contribution modeled across various mutation orders. By this means, Cerebra-Epistasis performs sequence and structure encoding only once to construct a reusable mutation-indexed atlas for a target protein. When evaluated across diverse multi-mutant benchmarks, Cerebra-Epistasis consistently achieves the leading performance in predicting both multi-mutant fitness and epistatic effects, with end-to-end training introducing further benefits to fitness prediction. The reusable mutation atlas also provides an interpretable decomposition of multi-mutant fitness and enables efficient evaluation of the large combinatorial fitness landscape.

Results

Overview of Cerebra-Epistasis

We propose Cerebra-Epistasis, a structure-aware framework that encodes the wild-type (WT) protein once and reuses the resulting representations to predict arbitrary-order mutant fitness (Fig 1A). The WT encoding is used to construct a mutation atlas covering all 20 amino-acid states at every sequence position. For a given multi-mutant, the corresponding entries are retrieved from the atlas and assembled to predict its fitness value.

The structure awareness is accomplished through Cerebra-Seq, a single-sequence structure predictor that combines ESMC [19] and ESM3 [20] representations through an mHC-based mixing

head [21], followed by Evoformer blocks and PSA [18] structure modules for information processing (Fig 1B). It predicts atomic-level protein structures and simultaneously generates residue-level node and residue-pair edge representations, providing structural context for downstream fitness prediction.

In the downstream fitness model, an SE(3)-equivariant network integrates the coordinates, node representations, and edge representations generated by Cerebra-Seq to capture the 3D geometry and residue-pair context, as well as ESM-2 [16] embedding to facilitate the residue-level representation. These features are transformed into a single-mutation score tensor **S** and an epistasis representation tensor **U** (Fig 1C), where **S** scores individual substitutions and **U** provides representations for modeling epistatic interactions. This design explicitly separates single-mutation contributions from the joint epistatic terms, effectively disentangling the components of multi-mutant fitness. The corresponding atlas entries can be assembled for arbitrary mutant combinations, enabling evaluation of the large combinatorial fitness landscape without repeated sequence or structure encoding.

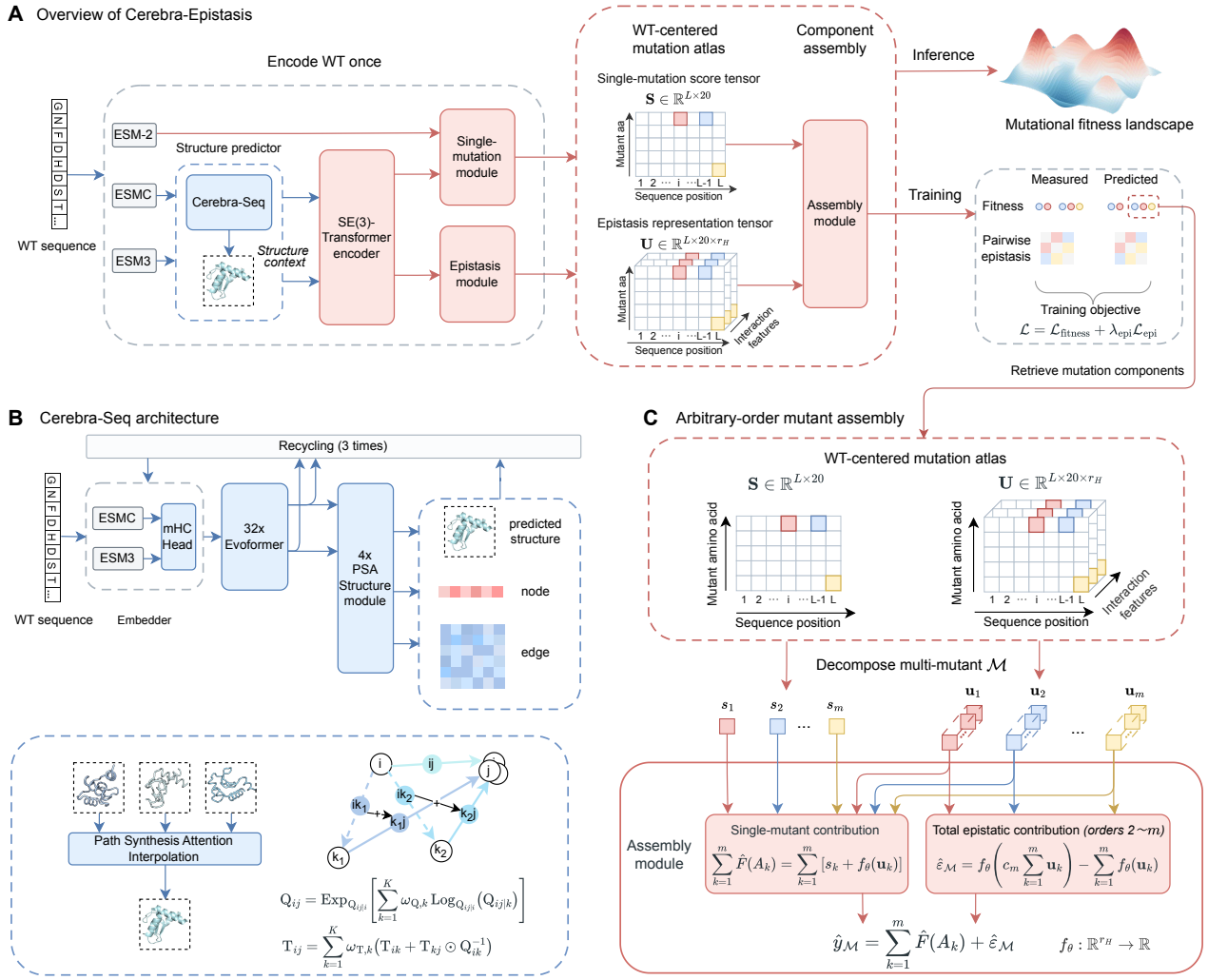


Figure 1. Overview of Cerebra-Epistasis. A, The WT protein is encoded once to construct a mutation-indexed atlas comprising the single-mutation score tensor **S** and epistasis representation tensor **U**. Mutation components are then assembled to predict fitness and epistasis across the mutant landscape. The framework supports downstream-only training with Cerebra-Seq frozen and end-to-end training. **B,** Architecture of Cerebra-Seq, which combines ESMC and ESM3 representations through an mHC-based mixing head, Evoformer blocks, and PSA structure modules to generate structural, node, and edge representations. **C,** Arbitrary-order mutant assembly through direct retrieval and permutation-invariant combination of the corresponding components from **S** and **U**, enabling lightweight fitness-landscape prediction without repeated sequence or structure encoding.

Cerebra-Seq achieves accurate single-sequence structure prediction

We evaluated Cerebra-Seq on two CAMEO [22] benchmarks to assess whether its single-sequence structure-prediction performance was reproducible across independent test periods. The cleaned CAMEO 2024 and CAMEO 2025 sets contain 642 and 573 protein targets released in the preceding two years, respectively. Cerebra-Seq was compared with ESMFold [16], OmegaFold [17], direct structure generation by ESM3 [20], and SPIRED [14]. Within each benchmark, all summary statistics and paired comparisons were calculated on the targets shared by all five methods. Prediction accuracy was evaluated using TM-score, which measures global fold similarity.

On CAMEO 2024, Cerebra-Seq achieved a mean TM-score of 0.83, second only to ESMFold at 0.86 (Fig 2A), presenting a mean paired advantage of 0.01, 0.04, and 0.05 over OmegaFold, ESM3, and SPIRED, respectively. In terms of the individual proteins, Cerebra-Seq outperformed ESMFold on 255 out of the 642 shared targets, which confirms its complementarity to this state-of-the-art baseline.

On CAMEO 2025, Cerebra-Seq again ranked second by the mean TM-score, reaching 0.81 compared with 0.83 for ESMFold (Fig 2B). Its advantage over OmegaFold, ESM3, and SPIRED was 0.02, 0.06, and 0.07, respectively (Fig 2C). At the individual-protein level, Cerebra-Seq exceeded ESMFold on 225 out of the 573 shared targets.

Across both benchmarks, Cerebra-Seq consistently outperformed direct ESM3 structure generation. This improvement suggests that integrating ESM3 with ESMC and subsequent structural refinement provides additional information for structure prediction. Although ESMFold achieved the highest mean TM-score, Cerebra-Seq outperformed it on substantial subsets of targets. The two CAMEO benchmarks show that Cerebra-Seq can predict all-atom protein structures from single sequences with sufficient accuracy.

Training and inference cost of Cerebra-Seq

The trainable structure-prediction component of Cerebra-Seq contains 77.0M parameters, and final-model training required 572 GPU-days. Using the same structure-component counting convention, the corresponding values reported in SPIRED-Fitness are 125M parameters and 85 GPU-days for SPIRED, 124M and 3,456 GPU-days for OmegaFold, and 690M and approximately 896 GPU-days for ESMFold [14]. Cerebra-Seq therefore has the smallest reported structure-predictor parameter count and a training cost below those of ESMFold and OmegaFold, although GPU-days are not directly comparable across different training protocols and hardware.

We benchmarked Cerebra-Seq against ESMFold, OmegaFold, and SPIRED on sequences of 100–1,000 residues for inference cost (Fig 2D). Albeit slower than ESMFold and OmegaFold at 100 residues (1.77 s vs. 0.95 and 1.23 s), Cerebra-Seq was the fastest predictor at every tested length from 200 residues onward. Across the 200–1,000-residue range, Cerebra-Seq took 20.20 s per target on average, corresponding to 1.84-, 4.53-, and 1.94-fold speedups over ESMFold, OmegaFold, and SPIRED, respectively. For sequences 1,000 residues long, the speedup reached 1.90-, 4.94-, and 1.55-fold, respectively.

These inference benchmarks further show that Cerebra-Seq attains structural accuracy and efficient single-sequence inference simultaneously, supporting its use as the structure-prediction module within Cerebra-Epistasis.

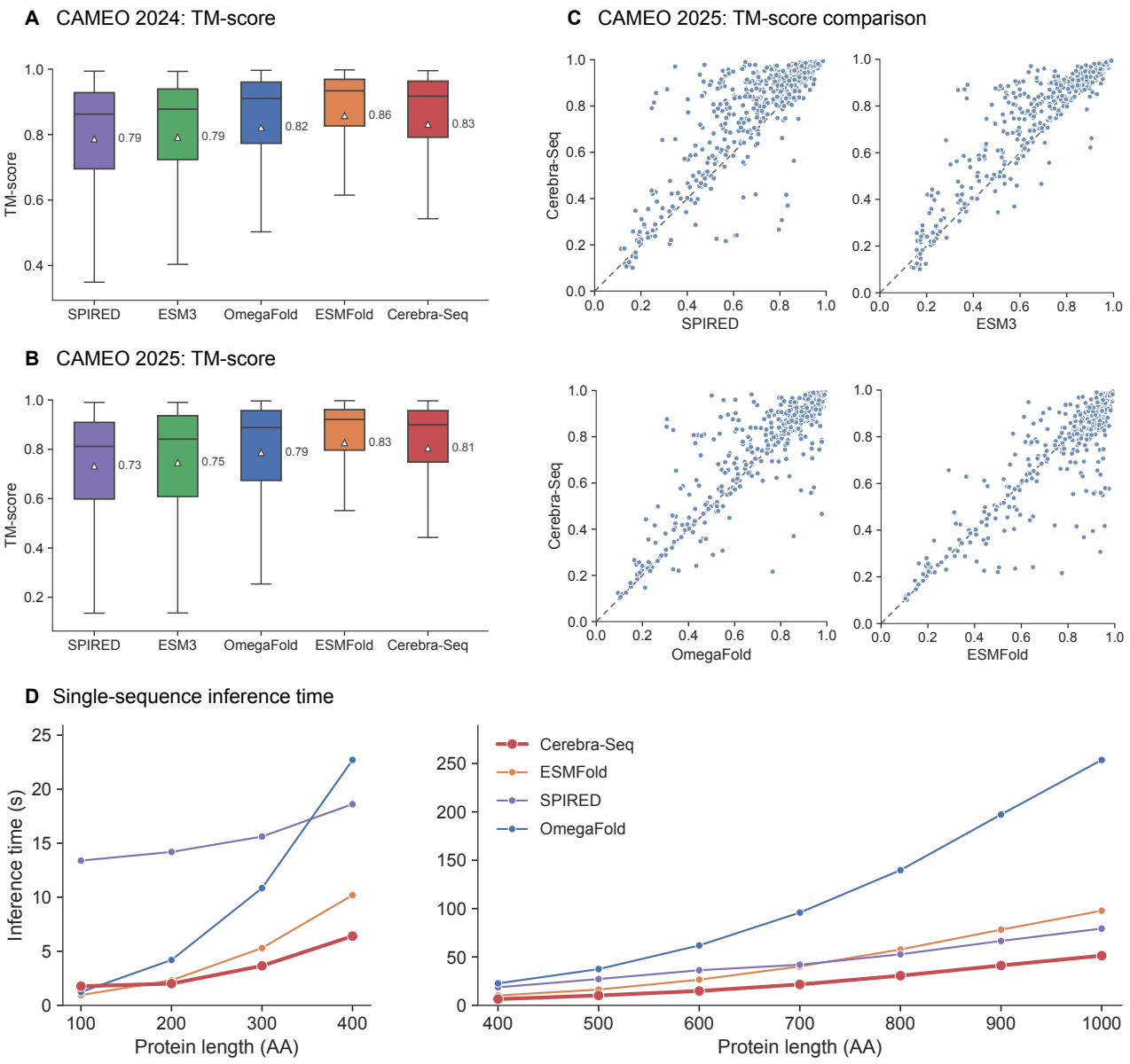


Figure 2. Accuracy and inference efficiency of Cerebra-Seq for single-sequence structure prediction. A,B, Per-target TM-score distributions for Cerebra-Seq and the baseline methods on the CAMEO 2024 (**A**) and CAMEO 2025 (**B**) benchmarks. White triangles and adjacent values indicate mean TM-scores; center lines denote medians, boxes represent the interquartile range, and whiskers extend to 1.5 times the interquartile range. **C,** Paired per-target comparison of Cerebra-Seq with SPIRED, direct ESM3 structure generation, OmegaFold, and ESMFold on CAMEO 2025. Each point denotes a target shared by all five methods. The diagonal line marks equal TM-scores; points above the line favor Cerebra-Seq. **D,** Single-sequence inference times of Cerebra-Seq, ESMFold, SPIRED, and OmegaFold for proteins ranging from 100 to 1,000 residues. The left and right panels show the 100–400- and 400–1,000-residue ranges, respectively.

Cerebra-Epistasis accurately predicts mutant fitness across diverse tasks

We evaluated Cerebra-Epistasis in two benchmark settings grounded in established protein-fitness evaluation protocols (Fig 3A,F). The 1M+ to 1M+ benchmark follows the random cross-validation setting commonly used in ProteinGym [5], with position balancing to ensure comparable mutation-site coverage across folds (Supplementary Section S2.3). The more protein-engineering-relevant Low-N 1M/2M to 3M+ benchmark follows the low-data higher-order extrapolation setting used in MULTI-evolve [9], by training on fewer than 300 single- and double-mutant measurements and testing on unseen higher-order combinations (Supplementary Section S2.4).

On the 1M+ to 1M+ benchmark, Cerebra-Epistasis achieved the strongest overall assay-level

performance (Supplementary Table S5), with statistics of assay-level results shown in Fig 3B,C. Specifically, it achieved a mean Spearman correlation of 0.85 and an NDCG (top 10%) of 0.95, compared with 0.84 and 0.94, respectively, for the second-best baseline ProteinNPT [7]. Notably, Cerebra-Epistasis slightly outperformed this leading supervised multi-mutant fitness predictor despite encoding only the wild-type protein rather than each mutant sequence, indicating that multi-mutant fitness can be effectively predicted within a shared wild-type context by learning how mutation-specific effects combine. This advantage was broadly maintained across mutation orders (Fig 3D,E).

Because practical protein engineering often relies on limited measurements of low-order mutants to identify promising higher-order variants, we next evaluated Cerebra-Epistasis in the Low-N 1M/2M to 3M+ benchmark, where it showed a clearer advantage over competing methods (Fig 3G,H; Supplementary Table S6). Specifically, it achieved an overall mean Spearman correlation of 0.65, compared with 0.62 for Kermut, 0.61 for MULTI-evolve, and 0.56 for ProteinNPT. When performance was averaged across mutation orders, Cerebra-Epistasis remained the strongest at 0.65, superior to 0.62 for Kermut, 0.63 for MULTI-evolve, and 0.57 for ProteinNPT. Whereas MULTI-evolve [9] learns epistatic patterns through a nonlinear fitness predictor trained on single- and double-mutant measurements, Cerebra-Epistasis uses a dedicated epistasis representation optimized with both fitness and epistasis supervision. As mutation order increased, all methods showed declining performance, but Cerebra-Epistasis declined more slowly, resulting in a progressively larger advantage at higher orders (Fig 3I-K). This trend supports the value of explicitly modeling epistasis for extrapolating from sparse low-order measurements to unseen higher-order combinations.

We also evaluated Cerebra-Epistasis in three additional benchmark settings. The 1M/2M to 3M+ benchmark further examined the model's ability to extrapolate from low-order measurements to unseen higher-order variants in the presence of more available data (Supplementary Section S2.5; Supplementary Table S7), while the cDNA proteolysis multi-mutant benchmark evaluated performance across a large collection of independent multi-mutant assays (Supplementary Section S2.6; Supplementary Table S8). We further validated Cerebra-Epistasis on an independent avGFP wet-lab–measured test set (Supplementary Section S2.7; Supplementary Table S9). Across all three additional benchmarks, Cerebra-Epistasis achieved the strongest overall performance among the tested methods.

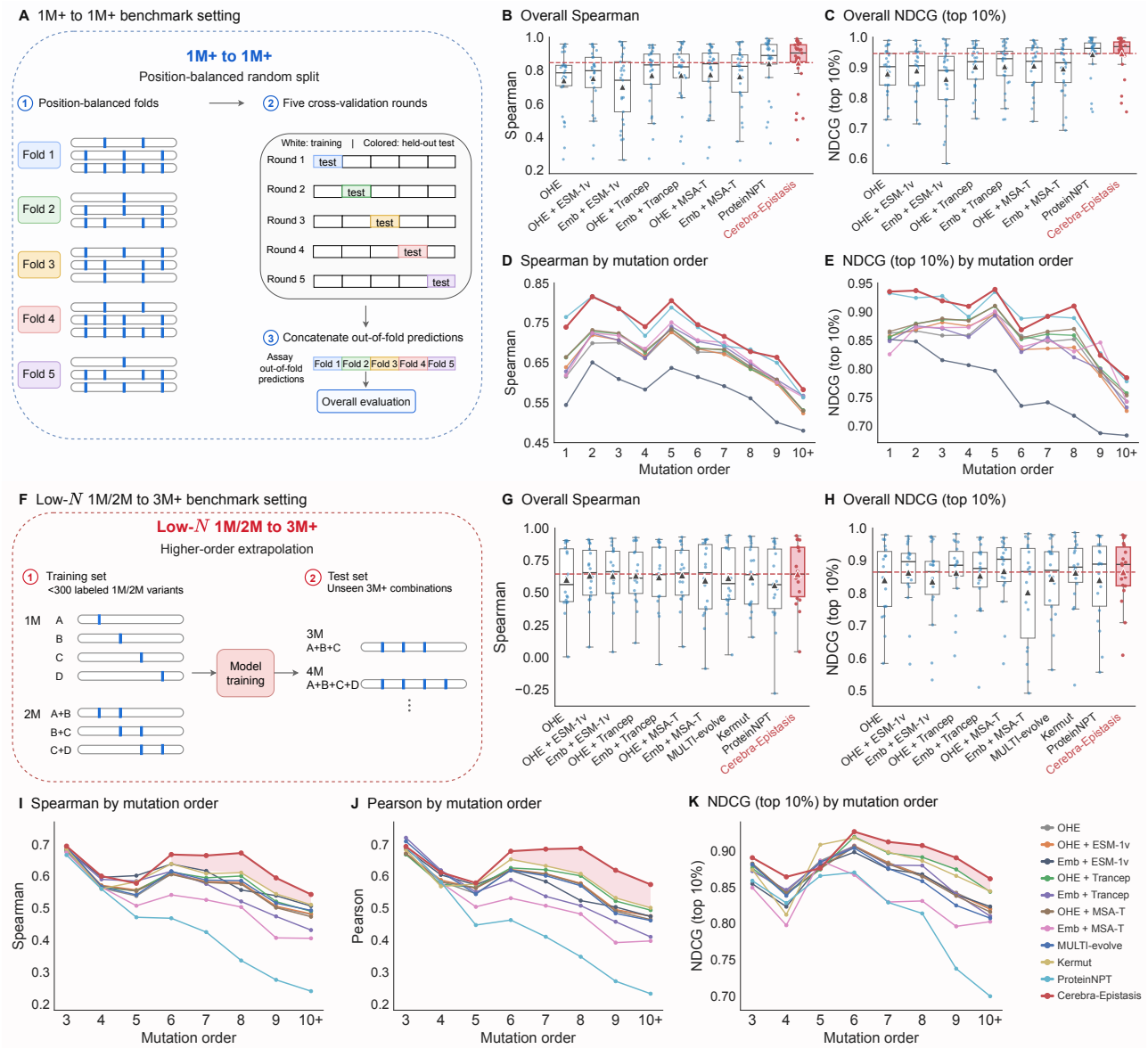


Figure 3. Cerebra-Epistasis accurately predicts mutant fitness across diverse tasks. A, Position-balanced 1M+ to 1M+ benchmark setting. **B,C,** Overall assay-level Spearman correlation and NDCG (top 10%), respectively; points denote assays, triangles denote assay means, and the red dashed line indicates the Cerebra-Epistasis mean. **D,E,** Assay-averaged Spearman correlation and NDCG (top 10%) at each mutation order, respectively. **F,** Low-N 1M/2M to 3M+ benchmark setting. **G,H,** Overall assay-level Spearman correlation and NDCG (top 10%) in the Low-N benchmark, respectively; points denote assays, triangles denote assay means, and the red dashed line indicates the Cerebra-Epistasis mean. **I–K,** Assay-averaged Spearman correlation, Pearson correlation, and NDCG (top 10%) at each mutation order in the Low-N benchmark, respectively.

Cerebra-Epistasis predicts unseen epistatic interactions and reveals biologically meaningful patterns

Epistasis—the non-additive interactions between mutations—is a major source of complexity in the protein fitness landscape. Unlike many variant-level fitness predictors, in which epistasis remains implicit and must be derived post hoc from predicted mutant fitness values, Cerebra-Epistasis explicitly separates epistatic effects from single-mutation contributions. To assess the contribution of explicit epistasis modeling, we compared Cerebra-Epistasis with two additive formulations, Ridge-additive and Cerebra-Epistasis-additive. Ridge-additive uses L_2-regularized linear regression to model multi-mutant fitness as the sum of single-mutation effects, whereas Cerebra-Epistasis-additive disables the

epistasis branch of Cerebra-Epistasis and retains only its single-mutation contributions. On the Low-N 1M/2M-to-3M+ benchmark, Cerebra-Epistasis consistently outperformed both additive formulations across mutation orders (Fig 4A).

To directly assess whether Cerebra-Epistasis could predict epistatic interactions in unseen mutation combinations, we designed a Low-N 1M/2M to held-out 2M (epistasis) benchmark. The constituent single mutations of each test double mutant were observed during training, while their specific combination was held out, thereby isolating prediction of unseen pairwise epistasis (Fig 4B; Supplementary Section S2.8). Cerebra-Epistasis substantially outperformed all applicable baselines across correlation- and ranking-based metrics (Fig 4C,D; Supplementary Table S10). It achieved a mean Spearman correlation of 0.55, compared with 0.35 for Kermut and 0.19 for MULTI-evolve, and a top-10% recall of 0.47, compared with 0.34 for Kermut and 0.33 for MULTI-evolve. Kermut [13] remained relatively competitive among the baselines, suggesting that 3D structural context is informative for modeling mutational coupling. Cerebra-Epistasis achieved further gains by explicitly modeling epistatic effects.

We next examined whether Cerebra-Epistasis could recover experimentally observed and biologically meaningful epistatic patterns across diverse protein systems. TrpB, the β-subunit of tryptophan synthase, catalyzes the final step of tryptophan biosynthesis. Johnston et al. [23] mapped a complete combinatorial fitness landscape across four active-site residues, including all double-, triple-, and quadruple mutants, enabling pairwise and higher-order epistasis to be quantified (Fig 4H). Across this landscape, Cerebra-Epistasis recovered epistatic effects from two- to four-body interactions and achieved the strongest performance across epistasis orders (Fig 4I). At the pairwise level, our reanalysis of the experimental landscape identified V227K/S228G as the strongest positive epistatic interaction among 2,166 double mutants (Fig 4F). Cerebra-Epistasis likewise ranked this pair first and reproduced the V227×S228 epistatic landscape across substitutions (Fig 4F,G). Structural mapping further showed that V227 and S228 are spatially adjacent near the PLP-bound active site (Fig 4E).

Beyond individual epistatic interactions, we asked whether Cerebra-Epistasis could recover broader structural patterns of epistasis using the GB1 landscape. GB1 is a small domain of streptococcal protein G that binds the Fc region of IgG. Olson et al. [24] comprehensively mapped pairwise epistasis across the GB1 domain and analyzed its relationship with protein structure. Following Olson et al.'s G9-centered structural analysis, we applied the same analysis to the measured and predicted epistatic interactions and observed similar spatial patterns (Fig 4J). More broadly, residue pairs with stronger measured epistasis tended to occur at shorter spatial distances, with 7 of the 10 strongest pairs forming structural contacts. The predicted interactions showed similar distance dependence and contact enrichment (Fig 4K,L).

Finally, we asked whether Cerebra-Epistasis could recover an experimentally established pattern of background-dependent epistasis in GB1. Wu et al. [6] found that higher-order interactions were enriched among sites 39, 41, and 54, whose side chains are spatially coupled in the GB1 core (Fig 4M), and further showed that the interaction between G41 and V54 depends on the amino acid at site 39. Our reanalysis showed that the G41–V54 interaction exhibited positive epistasis in the V39I, V39L, and V39M backgrounds, but negative epistasis in the V39V and V39W backgrounds. Cerebra-Epistasis closely reproduced this background-dependent pattern across site-39 substitutions, whereas ProteinNPT captured it less consistently (Fig 4N).



Figure 4. Cerebra-Epistasis predicts unseen epistasis and recovers biologically relevant interaction patterns. **A,** Cerebra-Epistasis, Cerebra-Epistasis-additive, and Ridge-additive on the Low-*N* 1M/2M-to-3M+ benchmark across mutation orders. **B,** Low-*N* 1M/2M-to-held-out-2M epistasis benchmark; all constituent single mutants remain in training while double-mutant combinations are held out in five position-balanced folds. **C,** Assay-level Spearman correlations for held-out pairwise epistasis prediction. **D,** Assay-level NDCG@10% and top recall@10%. Points denote assays, triangles denote assay means, and red dashed lines indicate the Cerebra-Epistasis means. **E,** Structural context of V227 and S228 in TrpB near the PLP-containing active site. **F,** Measured and predicted epistasis of 2,166 TrpB double mutants ordered by measured epistasis; V227K/S228G ranks first in both. **G,** Measured and predicted V227×S228 epistasis across amino-acid substitutions. **H,** Numbers of measured TrpB variants used to derive two-, three-, and four-body epistasis. **I,** Prediction of two-, three-, and four-body epistasis on the TrpB landscape. **J,** G9-centered epistasis mapped onto the GB1 structure. For each partner residue, epistasis was averaged across all available substitution pairs with G9 and centered across partner residues before structural mapping. **K,** Mean absolute epistasis for GB1 residue pairs versus spatial distance; blue points denote the 20 strongest measured pairs and the dashed line marks the 8-Å contact cutoff. **L,** Structural-contact enrichment among residue pairs ranked by mean absolute epistasis. The dashed line denotes the random expectation. **M,** GB1 wild-type structure highlighting sites V39, G41, and V54. **N,** Background-dependent G41–V54 epistasis. For each amino acid at site 39, conditional epistasis was calculated for all available G41/V54 substitution pairs and averaged. Shaded bands indicate the standard error of the mean (SEM).

Cerebra-Epistasis enables one-shot fitness landscape prediction

Protein engineering often involves screening a large number of variants. The number of possible variants increases exponentially with mutational depth. For a protein of only 100 residues, the sequence space already contains approximately 1.86×10^{14} possible five-mutation variants and 1.06×10^{26} possible ten-mutation variants (Fig 5B). Such a combinatorial growth trend highlights the importance of efficiency for computational baselines. Conventional supervised predictors require candidate-dependent feature construction and model evaluation for each variant. In contrast, Cerebra-Epistasis encodes the WT protein only once to generate the single-mutation score tensor S and epistasis representation tensor U. The entire fitness landscape can then be scored through lightweight indexing and assembly of the corresponding mutation components, without repeated sequence or structure encoding (Fig 5A).

To quantify this computational advantage, we compared Cerebra-Epistasis with two of the strongest supervised fitness predictors, ProteinNPT and Kermut. For training, we adopted two modes: i) varying the protein length while sampling a fixed number of 5,000 labeled variants; and ii) varying the training-set size at a fixed protein length of 300 residues. During inference, we relaxed the number of candidate variants to 50,000 in mode (i), while varying the candidate-library size at a fixed length of 300 residues in mode (ii) (Fig 5D,E). Clearly, Cerebra-Epistasis required substantially less training time as protein length and training-set size increased, while its inference time remained low across sequence lengths and nearly constant as the candidate library grew. The computational gap widened further when screening complete higher-order landscapes. Based on the measured inference throughput, we estimated by extrapolation that screening the complete 3-mutant landscape of a length-100 protein would require only 3.70 min with Cerebra-Epistasis, compared with 2.57 days for Kermut and 2.60 years for ProteinNPT; for the complete 4-mutant landscape, the corresponding time was 1.18 days, 3.24 years, and 1.20×10^3 years, respectively. These estimates correspond to an approximately 99.9% reduction in inference time relative to Kermut and a greater than 99.999% reduction relative to ProteinNPT at both mutation orders (Fig 5C).

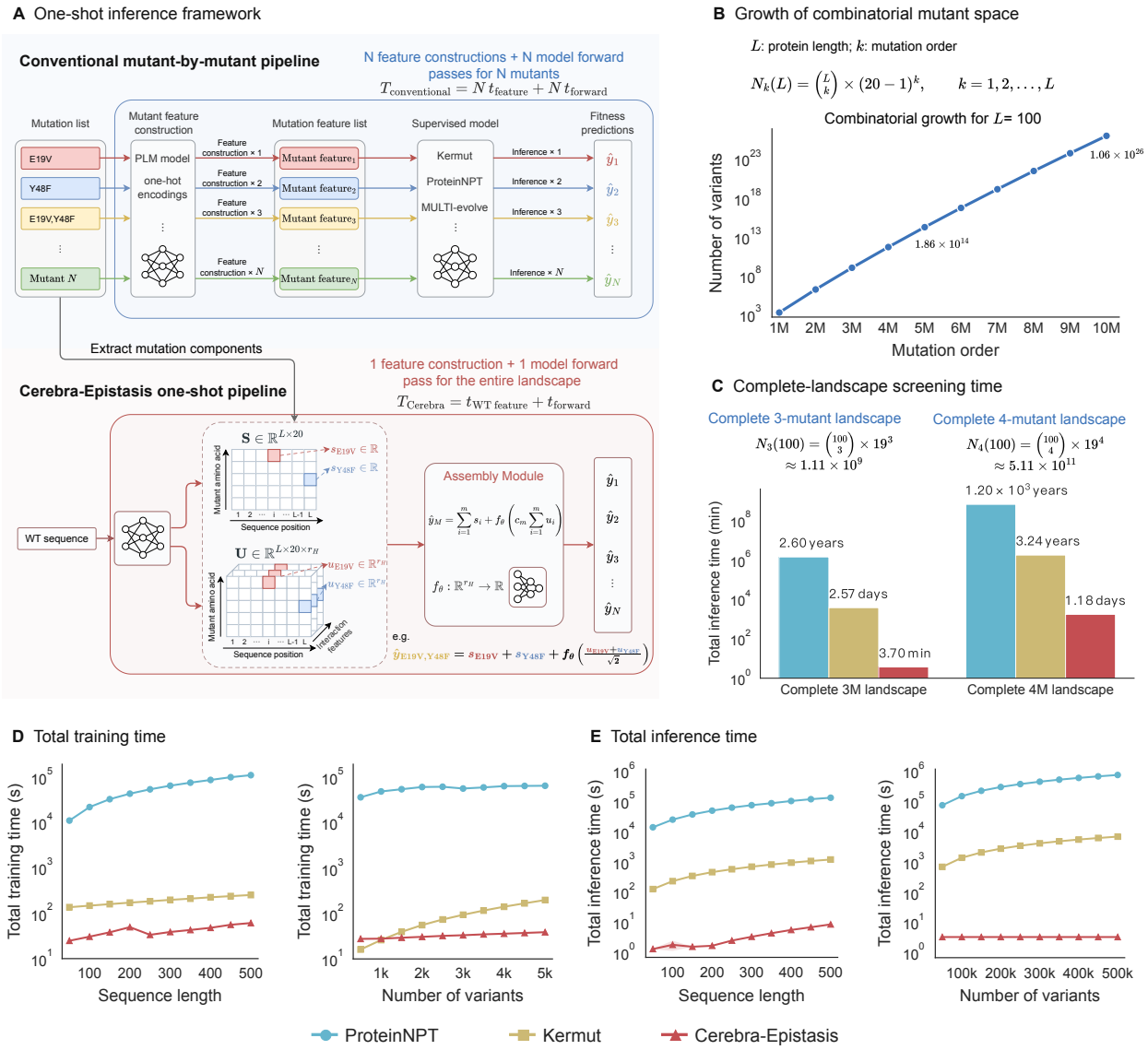


Figure 5. One-shot inference for fitness landscape prediction. **A,** Comparison of conventional mutant-by-mutant prediction with the one-shot inference paradigm of Cerebra-Epistasis. **B,** Combinatorial growth of mutant space for a protein of length $L = 100$. **C,** Estimated inference time for screening complete 3M and 4M landscapes at $L = 100$, extrapolated from measured inference throughput. **D,** Overall training time, including feature generation and model training, with increasing sequence length at 5,000 labeled variants (left) and increasing training-set size at $L = 300$ (right). **E,** Overall inference time, including feature generation and prediction, with increasing sequence length for 50,000 candidates (left) and increasing candidate-library size at $L = 300$ (right).

End-to-end training improves fitness prediction while preserving structural accuracy

Backpropagation through protein structure predictors incurs substantial computational and memory costs. Consequently, structure predictors and downstream fitness models are typically optimized separately. The lightweight Cerebra-Seq architecture makes end-to-end training with Cerebra-Epistasis computationally tractable. We therefore asked whether task-specific adaptation could improve fitness prediction without degrading the structural representations acquired during pretraining.

Reference proteins from the cDNA proteolysis multi-mutant dataset were clustered by sequence similarity and assigned, together with all associated variants, to five folds. In each round, end-to-end-trained and frozen-Cerebra-Seq models were trained on four folds and evaluated on the held-out

fold using identical partitions. Further details of the protein-level data split and training protocol are provided in Supplementary Section S3.5.

The two settings exhibited distinct training trajectories (Fig 6A). End-to-end training maintained higher five-fold mean Spearman correlations after early crossings of the training trajectories. Under the reported 100-epoch schedule, the end-to-end-trained model achieved a five-fold mean Spearman correlation of 0.72 ± 0.02, compared with 0.68 ± 0.02 for frozen Cerebra-Seq. Frozen Cerebra-Seq peaked at 0.71 at epoch 36, whereas the end-to-end-trained model reached its maximum of 0.73 at epoch 54. These trajectories support a sustained benefit of end-to-end training for fitness prediction on held-out cDNA proteins during later training, and indicate alleviation of late-training overfitting.

To assess whether this pattern extended beyond the cDNA proteins, we trained both settings on the complete dataset and evaluated them on the independent S461 $\Delta\Delta G$ benchmark (Fig 6B; Supplementary Table S11). End-to-end training increased mean Spearman, Kendall and Pearson correlations from 0.61, 0.48 and 0.66 to 0.71, 0.57 and 0.71, respectively. Similarly, the Top-5 and Top-10 precision rose from 0.52 and 0.67 to 0.57 and 0.74, respectively. The end-to-end-trained model matched the highest mean Spearman and Kendall correlations and achieved the highest mean Pearson correlation and Top-10 precision, although its mean Top-5 precision remained below the benchmark-leading value of 0.63. These results extend the cross-validation observation by showing improved performance after full-dataset training, when evaluated on an external protein-stability benchmark.

We next asked whether these fitness gains came at the cost of the pretrained structural representation. Structural performance was evaluated on held-out proteins with experimentally determined structures, with epoch 0 corresponding to the pretrained baseline before end-to-end training (Fig 6C,D). Mean side-chain χ torsion-angle loss decreased from 0.45 ± 0.02 to 0.44 ± 0.01 by epoch 100, corresponding to a reduction of 2.60 %. Mean lDDT remained between 0.79 and 0.81 and ended at 0.80 ± 0.02, comparable to the baseline range of 0.80 ± 0.03. Ablations showed that fitness gradients improved prediction beyond structural supervision alone, while additional structural supervision modestly improved fitness prediction and helped preserve structural accuracy (Supplementary Fig S5). Taken together, under the reported schedules, end-to-end training yielded higher final-epoch cDNA and S461 performance while modestly improving the side-chain torsion sampling and largely preserving the local structural quality.

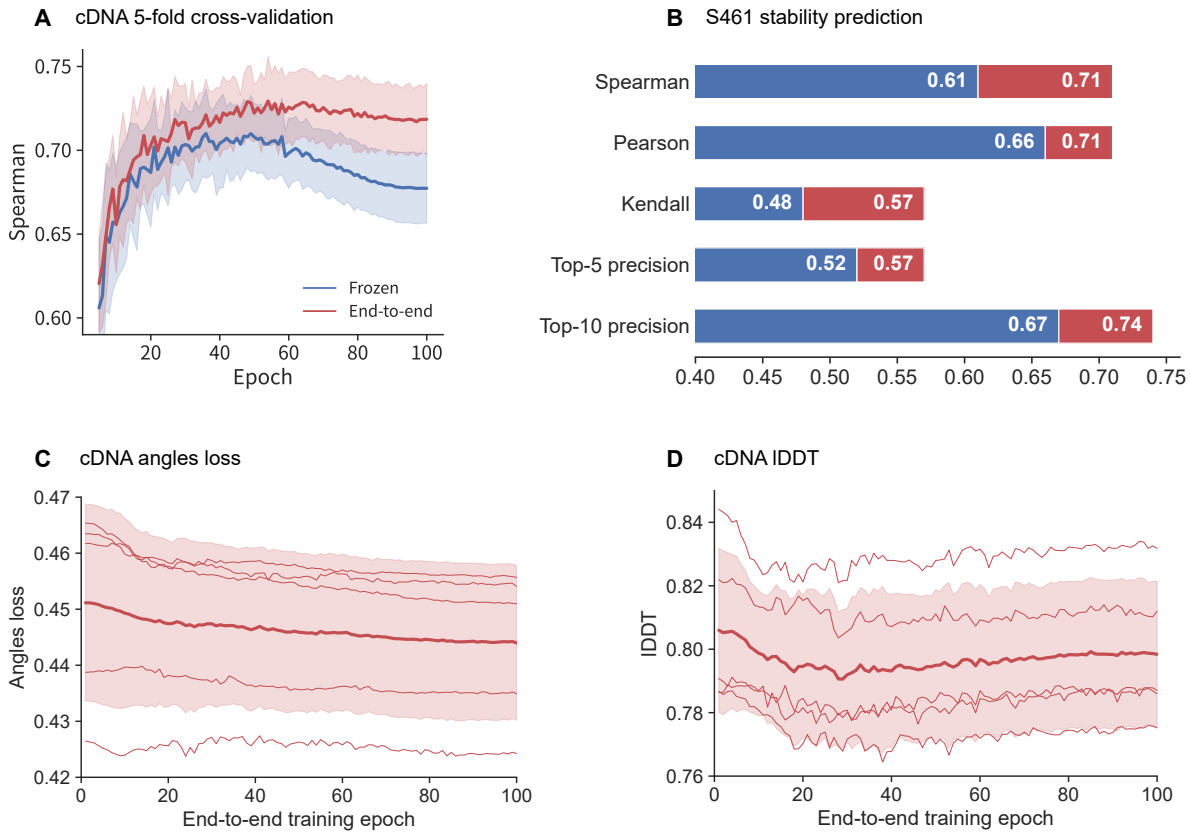


Figure 6. Effect of end-to-end training on downstream fitness prediction and structure prediction. A, Spearman correlations on held-out cDNA proteins in five-fold cross-validation, with splits defined by sequence-similarity clusters. Blue curves show the frozen-Cerebra-Seq baseline, in which only the downstream model was trained; red curves show end-to-end training of both components. Both settings used identical partitions. Curves show the five-fold mean at shared evaluation epochs, and shading indicates ± one standard deviation. **B,** Prediction performance on the independent S461 stability benchmark after training on the complete cDNA proteolysis dataset. Blue bars show the frozen-Cerebra-Seq baseline, and red extensions indicate the improvement from end-to-end training. Numeric labels report the mean scores across proteins for both settings. **C,** Side-chain χ torsion-angle loss during end-to-end training; lower values indicate better predictions. **D,** lDDT of the predicted structures during end-to-end training. In **C** and **D**, metrics were computed on proteins with experimental structures in each held-out fold. Epoch 0 denotes the pretrained Cerebra-Seq baseline prior to end-to-end training. Thin red curves represent individual folds, the thick red curve shows the five-fold mean, and red shading indicates ± one standard deviation.

Discussion

Accurate construction of the multi-mutant fitness landscape requires epistasis to be modeled across varying mutation orders without increasing landscape-scale inference cost. Cerebra-Epistasis addresses these linked requirements through a reusable mutation-indexed atlas. The assembly module retrieves the relevant components to predict each mutant's fitness and its epistatic contributions. Across the evaluated benchmark settings, Cerebra-Epistasis improves the overall multi-mutant fitness prediction over leading baselines. Its growing advantage at increasing mutation orders further supports the importance of explicit epistasis modeling. When the same mutation-specific representations are assembled for different mutation sets, the predicted epistasis is supposed to reflect the influence of mutational background. This design bridges low-order interaction expansions and nonlinear variant-level predictors while avoiding repeated sequence and structure encoding for every candidate [6, 7, 9].

Earlier sequence–structure models show that structural context can improve variant-effect predic-

tion [11, 12, 14]. Our structural-feature ablations extend this evidence to higher-order multi-mutant fitness extrapolation. Adding Cerebra-Seq coordinates, residue-level node representations and residue-pair edge representations to sequence features generally benefits prediction. End-to-end training of Cerebra-Seq and the fitness model further improves fitness prediction while largely preserving local structural accuracy, as measured by lDDT. These findings support end-to-end training of the structure predictor and fitness model rather than treating predicted structure solely as fixed preprocessing. However, Cerebra-Seq still requires intensive pretraining to maintain its competitive predictive power. Improving structural accuracy without sacrificing end-to-end trainability therefore remains an important development goal.

The held-out interaction benchmark and analyses of TrpB and GB1 suggest that the explicit epistasis output captures patterns beyond aggregate fitness. These examples include local structural coupling, enrichment of spatial contacts, and changes in interaction sign across sequence backgrounds. However, the WT-centered formulation places an important boundary on the mechanistic interpretation of these patterns. The mutation-indexed atlas is constructed from the WT sequence and its predicted structure. It therefore does not explicitly represent mutation-induced rearrangements, conformational ensembles, or changes in oligomeric state, ligands or binding partners. This limitation may matter most when mutations alter structural states or context-dependent molecular interactions. Future models could incorporate conformational ensembles, molecular context, or selective mutant-specific structural refinement, while retaining most of the computational benefit of atlas reuse.

One-time WT encoding shifts the expensive sequence and structure computation from large candidate libraries to a single reference protein. The assembly module then retrieves and combines the relevant atlas components for each candidate. This inference paradigm underlies the projected speed advantage over models that encode and evaluate each mutant separately. However, efficient assembly does not eliminate the combinatorial growth of the candidate space. Likewise, the dependence on double-mutant coverage shows that explicit epistasis modeling cannot replace informative experimental measurements. Uncertainty-calibrated ranking and active selection could focus computation and experiments on the most informative regions of the landscape. Prospective design cycles will be needed to determine whether fitness and epistasis predictions improve protein optimization beyond retrospective benchmarks. Within these limits, Cerebra-Epistasis provides a unified framework for structure-aware multi-mutant fitness prediction, explicit epistasis modeling, and efficient landscape-scale inference.

Methods

Cerebra-Epistasis links single-sequence structure prediction with protein fitness landscape modeling through a composable, epistasis-aware mutation atlas (Fig 1). For a given target protein, Cerebra-Seq first predicts residue-level and pairwise representations together with side-chain-resolved coordinates. These outputs are encoded once by an SE(3)-Transformer and converted into two mutation-indexed tensors covering all 20 amino-acid states at every sequence position. An assembly module then retrieves the entries corresponding to a specified mutation set and predicts both its fitness and its collective epistatic contribution. This design provides explicit 3D context while avoiding repeated sequence and structure encoding for individual mutants.

Cerebra-Seq architecture

Single-sequence feature construction

Cerebra-Seq was developed from our previously reported MSA-based Cerebra architecture [18]. The original model integrates evolutionary information from an MSA and reconciles parallel coordinate hypotheses through path synthesis attention (PSA). Cerebra-Seq retains the Evoformer and PSA-based structure-generation principles but removes the requirement for an MSA. Instead, frozen ESMC and ESM3 models provide complementary per-residue embeddings directly from the target sequence. Their concatenated 2,688-dimensional representation is projected to 2,560 channels and reshaped into ten streams of width 256.

To enable information exchange among these streams while preserving stable signal propagation, we introduced a stream-mixing head inspired by manifold-constrained hyper-connections (mHC) [21]. A learnable 10×10 matrix is projected by Sinkhorn normalization onto the space of doubly stochastic matrices and used to mix the streams:

$$\widetilde{\mathbf{Y}}'_{i,t,:} = \sum_{s=1}^{10} H_{st} \widetilde{\mathbf{Y}}_{i,s,:}, \qquad \mathbf{H1} = \mathbf{1}, \quad \mathbf{H}^{\mathsf{T}}\mathbf{1} = \mathbf{1}. \tag{1}$$

This module adapts the residual-stream mixing principle of mHC to protein language-model feature fusion rather than implementing the complete mHC architecture. Learnable amino-acid-type embeddings are added to the mixed streams to form the initial single representation. In parallel, relative sequence positions and the amino-acid identities of each residue pair initialize the pair representation. The single and pair representations are then jointly refined by the Evoformer trunk with recycling.

Structure generation and downstream representations

The refined representations are supplied to a PSA-based structure module derived from Cerebra. For multiple approximately uniformly distributed anchor residues, the module predicts residue translations and rotations in the corresponding local coordinate systems. PSA combines direct coordinate predictions with transformations propagated through intermediate anchors, integrating multiple internally consistent structural hypotheses [18]. Predicted residue frames and torsion angles are subsequently converted into side-chain-resolved atom14 coordinates.

Cerebra-Seq outputs a residue representation $\mathbf{N} \in \mathbb{R}^{L\times 256}$, a pair representation $\mathbf{E} \in \mathbb{R}^{L\times L\times 128}$, and atom14 coordinates $\mathbf{X} \in \mathbb{R}^{L\times 14\times 3}$ for the downstream fitness model. Unless otherwise specified, Cerebra-Seq is frozen during assay-specific fitness training. In end-to-end training experiments, fitness gradients are additionally propagated through Cerebra-Seq, allowing its structure-aware representations to adapt to mutation-effect prediction. Detailed architectural and implementation settings are provided in Supplementary Algorithm S2.

Downstream fitness model architecture

For a fitness landscape centered at the wild type, consider a mutant $\mathcal{M} = \{A_k\}_{k=1}^{m}$ containing m substitutions, where each substitution $A_k = (i_k, a_k)$ is specified by its sequence position i_k and mutant amino-acid state a_k. Its fitness can be decomposed into two components: the summed contributions of

its constituent single mutations and the combined contributions of epistatic interactions of different orders:

$$F(\mathcal{M}) = \sum_{k=1}^{m} F(A_k) + \sum_{1\le p<q\le m} \varepsilon^{(2)}_{A_pA_q} + \sum_{1\le p<q<r\le m} \varepsilon^{(3)}_{A_pA_qA_r} + \cdots + \varepsilon^{(m)}_{A_1\cdots A_m}. \tag{2}$$

Here, $F(A_k)$ denotes the fitness contribution of single mutation A_k relative to the wild type, whereas $\varepsilon^{(n)}$ denotes an n-body epistatic contribution. The first term therefore represents the total contribution of the constituent single mutations, while the remaining terms collectively represent the total epistatic contribution from pairwise to m-body interactions.

Motivated by this decomposition, Cerebra-Epistasis comprises two complementary branches: a single-mutation score branch that provides mutation-specific scalar scores and an epistasis representation branch that constructs mutation-specific representations for modeling collective epistatic contributions.

Fig 1A,C and Supplementary Algorithm S1 summarize the overall workflow, which consists of three stages. First, Cerebra-Seq-derived structure-aware features of the wild-type protein are processed once by the SE(3)-Transformer encoder. Second, the resulting structure-aware representations, together with the residue-level ESM-2 embedding, are transformed into a single-mutation score tensor $\mathbf{S} \in \mathbb{R}^{L\times 20}$ and an epistasis representation tensor $\mathbf{U} \in \mathbb{R}^{L\times 20\times r_H}$. Third, for an arbitrary-order mutant, the corresponding entries are indexed from $\mathbf{S}$ and $\mathbf{U}$ and assembled to predict its fitness and total epistatic contribution.

SE(3)-Transformer encoding of the wild-type protein

Cerebra-Epistasis is designed as a structure-aware fitness prediction model. To effectively integrate 3D structural information while preserving the geometric transformation properties of the input, we employ an SE(3)-equivariant neural network to encode the Cerebra-Seq-derived structure-aware features of the wild-type protein. For a protein of length L, these features comprise Cerebra-Seq-derived residue representations $\mathbf{N} \in \mathbb{R}^{L\times 256}$, pairwise representations $\mathbf{E} \in \mathbb{R}^{L\times L\times 128}$, and predicted atom14 coordinates $\mathbf{X} \in \mathbb{R}^{L\times 14\times 3}$. The encoder follows the general formulation of the SE(3)-Transformer [25].

For each central residue, the neighborhood includes spatially proximal residues together with residues exhibiting strong Cerebra pairwise representations. We introduce two task-specific adaptations to better integrate these features within the SE(3)-equivariant encoder. First, the input equivariant convolution constructs degree-1 geometric features from the atom14 coordinates of each neighboring residue expressed relative to the C_α coordinate of the central residue. This center-referenced representation preserves local atom-level geometry while removing dependence on global translation. Second, the projected Cerebra pairwise representations are combined with Fourier-encoded inter-residue distances to parameterize the radial components of the equivariant kernels, allowing SE(3)-equivariant message passing to incorporate both explicit geometry and learned residue-pair context.

The encoder outputs a degree-0 invariant representation $\mathbf{Z} \in \mathbb{R}^{L\times 128}$, a degree-1 equivariant representation $\mathbf{V} \in \mathbb{R}^{L\times 32\times 3}$, and local distance statistics $\mathbf{D}^{\text{stat}} \in \mathbb{R}^{L\times 2}$, comprising the mean and standard deviation of the C_α distances within the selected neighborhood of each residue. The wild-type protein is encoded only once, and the resulting representations are shared by the downstream mutation-specific branches. Detailed SE(3)-Transformer operations and implementation of the two adaptations are provided in Supplementary Algorithm S5, Supplementary Algorithm S14, and

Supplementary Algorithm S13.

Construction of the single-mutation score and epistasis representation tensors

The encoded wild-type representations are subsequently transformed into two mutation-specific tensors covering all 20 amino-acid states at every sequence position: a single-mutation score tensor $\mathbf{S} \in \mathbb{R}^{L\times 20}$ and an epistasis representation tensor $\mathbf{U} \in \mathbb{R}^{L\times 20\times r_H}$. The former provides a scalar score for each position–amino-acid combination, whereas the latter provides a mutation-specific representation that can be assembled with other substitutions to model epistatic effects.

Specifically, the single-mutation score branch maps the invariant representation $\mathbf{Z} \in \mathbb{R}^{L\times 128}$ to sequence–structure-informed substitution scores and combines them with a stochastically gated correction derived from the residue-level ESM-2 embedding. In parallel, the epistasis representation branch converts the equivariant representation $\mathbf{V} \in \mathbb{R}^{L\times 32\times 3}$ into channel-wise magnitudes $\|\mathbf{V}\| \in \mathbb{R}^{L\times 32}$ and concatenates them with $\mathbf{Z}$ and the local distance statistics $\mathbf{D}^{\text{stat}} \in \mathbb{R}^{L\times 2}$. The resulting residue-level features are transformed by a multilayer perceptron (MLP) into a position-specific context tensor $\mathbf{H} \in \mathbb{R}^{L\times r_H}$. A learnable amino-acid embedding dictionary $\mathbf{E}^{\text{AA}} \in \mathbb{R}^{20\times r_H}$ assigns an r_H-dimensional representation to each amino-acid state. The position-specific context and amino-acid embeddings are combined through broadcasted element-wise multiplication to produce an intermediate mutation-specific tensor $\widetilde{\mathbf{U}} \in \mathbb{R}^{L\times 20\times r_H}$, followed by a learnable linear mixing layer to obtain $\mathbf{U} \in \mathbb{R}^{L\times 20\times r_H}$.

Finally, both tensors are centered relative to the wild-type amino acid at each position by subtracting the corresponding wild-type score or representation from all amino-acid states, such that $\mathbf{S}_{i,w_i} = 0$ and $\mathbf{U}_{i,w_i,:} = \mathbf{0}$. The resulting tensors therefore encode mutation-induced changes relative to the wild-type state and can be directly indexed for arbitrary-order mutant assembly. Details of the two tensor constructions are provided in Supplementary Algorithm S6 through Supplementary Algorithm S8.

Assembly and prediction of arbitrary-order mutants

For a mutant $\mathcal{M} = \{A_k = (i_k, a_k)\}_{k=1}^{m}$, its mutation-specific score and epistasis representation are retrieved directly from the precomputed tensors:

$$s_k = \mathbf{S}_{i_k,a_k}, \qquad \mathbf{u}_k = \mathbf{U}_{i_k,a_k,:}. \tag{3}$$

The selected epistasis representations are assembled through a permutation-invariant summation, optionally scaled according to mutation order, and the mutant fitness is predicted as

$$\widehat{y}_{\mathcal{M}} = \sum_{k=1}^{m} s_k + f_\theta\left(c_m \sum_{k=1}^{m} \mathbf{u}_k\right), \tag{4}$$

where c_m denotes the mutation-order scaling factor, with $c_1 = 1$, and f_θ is a shared MLP that maps an epistasis representation to a scalar response.

To distinguish the contributions of individual substitutions from those arising through their joint assembly, the prediction can be decomposed into predicted single-mutation contributions and a total

epistatic contribution:

$$\widehat{y}_{\mathcal{M}} = \underbrace{\sum_{k=1}^{m} \left[s_k + f_\theta(\mathbf{u}_k) \right]}_{\sum_{k=1}^{m} \widehat{F}(A_k)} + \underbrace{\left[f_\theta\left(c_m \sum_{k=1}^{m} \mathbf{u}_k \right) - \sum_{k=1}^{m} f_\theta(\mathbf{u}_k) \right]}_{\widehat{\varepsilon}_{\mathcal{M}}}. \tag{5}$$

Here, $\widehat{F}(A_k) = s_k + f_\theta(\mathbf{u}_k)$ is the predicted single-mutation contribution of substitution A_k. Although $\mathbf{U}$ is learned to support interaction-aware assembly, $f_\theta(\mathbf{u}_k)$ also provides a mutation-specific unary contribution. The predicted total epistatic contribution $\widehat{\varepsilon}_{\mathcal{M}}$ is defined after subtracting these constituent unary responses from their joint response. Consequently, $\widehat{\varepsilon}_{\mathcal{M}} = 0$ for a single mutant by construction. For a double mutant, $\widehat{\varepsilon}_{\mathcal{M}}$ corresponds to predicted pairwise epistasis, whereas for higher-order mutants it represents the aggregate epistatic contribution beyond the sum of the constituent single-mutation contributions, consistent with the n-body fitness decomposition in Equation (2).

The permutation-invariant assembly and shared MLP enable Cerebra-Epistasis to model collective interactions among an arbitrary number of substitutions, independently of mutation order and without explicitly parameterizing individual pairwise or higher-order interaction terms. Once $\mathbf{S}$ and $\mathbf{U}$ have been constructed from a single encoding of the wild-type protein, large mutant libraries can be evaluated through tensor indexing and lightweight assembly without repeated sequence or structure encoding. Detailed assembly operations and the architecture of f_θ are provided in Supplementary Algorithm S9 and Supplementary Algorithm S10.

Cerebra-Epistasis training strategy

Cerebra-Seq was first pretrained independently for protein structure prediction using structural supervision. Cerebra-Epistasis was subsequently trained using experimentally measured protein fitness data. Unless otherwise specified, Cerebra-Seq was kept frozen during downstream training; end-to-end training updated both Cerebra-Seq and the downstream fitness model. Full training objectives and implementation details are provided in the Supplementary Section S3, and benchmark-specific datasets and data-splitting protocols are described in Supplementary Section S2.

Evaluation metrics

Metrics for structure prediction

We used TM-score [26] and lDDT [27] to evaluate the quality of predicted structures.

TM-score. TM-scores were calculated with the TM-score program [26]. TM-score measures the global topological similarity between a predicted structure and its reference after optimal alignment. The score is normalized by protein length and ranges from 0 to 1; values above 0.5 are generally considered to indicate the same overall fold.

Local distance difference test. The local distance difference test (lDDT) measures local structural accuracy without global superposition [27]. lDDT was computed over all Cα atom pairs separated by

less than 15 Å in the reference structure. For each pair, the absolute difference between the predicted and reference distances was compared against thresholds of 0.5, 1, 2 and 4 Å. The lDDT score is the mean, over the four thresholds, of the fraction of satisfied pairs.

Metrics for fitness prediction

For an assay containing n variants, let y_i and $\hat{y}_i$ denote the measured and predicted fitness values of variant i, respectively. Fitness and epistasis prediction performance was evaluated using Pearson correlation, Spearman correlation, Kendall's τ, NDCG (top 10%), and Top recall (top 10%). For the S461 protein stability benchmark, Top-5 and Top-10 precision were additionally used following the evaluation protocol of SPIRED-Stab [14]. For the avGFP experimental benchmark, NDCG (top 5) and Top recall (top 5) were used instead.

Pearson correlation. The Pearson correlation coefficient measures the linear relationship between measured and predicted values:

$$r = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - \bar{\hat{y}})^2}}, \tag{6}$$

where $\bar{y}$ and $\bar{\hat{y}}$ denote their respective means.

Spearman correlation. Spearman correlation measures the monotonic relationship between measured and predicted values and is calculated as the Pearson correlation between their ranks:

$$\rho = r\left(R(\mathbf{y}), R(\hat{\mathbf{y}})\right), \tag{7}$$

where $R(\cdot)$ denotes rank transformation.

Kendall's rank correlation. Kendall's τ measures ranking consistency based on concordant and discordant pairs. Kendall's τ_b, which accounts for ties, was used in all calculations.

NDCG. Following the ranking-based evaluation used for protein fitness prediction, measured fitness values were first converted to non-negative gains by min–max normalization:

$$g_i = \frac{y_i - \min_j y_j}{\max_j y_j - \min_j y_j}. \tag{8}$$

Let K denote the number of top-ranked variants considered, and let π_j denote the index of the variant at predicted rank j. The discounted cumulative gain over the top K predicted variants and its ideal value are defined as

$$\mathrm{DCG}_K = \sum_{j=1}^{K} \frac{g_{\pi_j}}{\log_2(j+1)}, \qquad \mathrm{IDCG}_K = \sum_{j=1}^{K} \frac{g_{\pi_j^*}}{\log_2(j+1)}, \tag{9}$$

where π_j^* denotes the ranking obtained by sorting variants according to measured fitness. NDCG is then defined as

$$\text{NDCG}_K = \frac{\text{DCG}_K}{\text{IDCG}_K}. \tag{10}$$

For the main fitness and epistasis benchmarks, $K = \lfloor 0.1n \rfloor$, corresponding to NDCG (top 10%). For the avGFP experimental benchmark, $K = 5$, corresponding to NDCG (top 5).

Top recall. Top recall measures the overlap between the top K variants according to experimental measurements and the top K variants according to model predictions. Let $\mathcal{T}_K$ and $\mathcal{P}_K$ denote these two sets, respectively. The metric is defined as

$$\text{Top recall}_K = \frac{|\mathcal{P}_K \cap \mathcal{T}_K|}{|\mathcal{T}_K|}. \tag{11}$$

For the main fitness and epistasis benchmarks, $K = \lfloor 0.1n \rfloor$, corresponding to Top recall (top 10%). For the avGFP experimental benchmark, $K = 5$, corresponding to Top recall (top 5).

Top-K precision. For the S461 benchmark, Top-K precision measures the fraction of the K highest-ranked predicted mutations that are also among the K mutations with the highest experimentally measured values. Let $\mathcal{P}_K$ and $\mathcal{T}_K$ denote the predicted and experimentally determined top-K sets, respectively. The metric is defined as

$$\text{Top-K precision} = \frac{|\mathcal{P}_K \cap \mathcal{T}_K|}{K}. \tag{12}$$

We used $K = 5$ and $K = 10$ for S461.

Data availability

The benchmark datasets, precomputed input features, Cerebra-Epistasis training code, and prediction outputs generated by Cerebra-Epistasis and the baseline methods are publicly available on Zenodo at https://doi.org/10.5281/zenodo.22899137.

Code availability

Cerebra-Seq and Cerebra-Epistasis were implemented in PyTorch. The source code and model weights for Cerebra-Seq are publicly available on Hugging Face at https://huggingface.co/GongLab-THU/Cerebra-Seq. The source code for Cerebra-Epistasis is publicly available on GitHub at https://github.com/xulab-research/Cerebra-Epistasis. A web server for online Cerebra-Epistasis prediction is available at https://structpred.life.tsinghua.edu.cn/server_cerebra_epistasis.html.

References

[1] Philip A. Romero and Frances H. Arnold. "Exploring protein fitness landscapes by directed evolution". In: *Nature Reviews Molecular Cell Biology* 10.12 (Dec. 2009), pp. 866–876. ISSN: 1471-0080. DOI: 10.1038/nrm2805. URL: http://dx.doi.org/10.1038/nrm2805.

[2] Douglas M Fowler and Stanley Fields. "Deep mutational scanning: a new style of protein science". In: *Nature Methods* 11.8 (July 2014), pp. 801–807. ISSN: 1548-7105. DOI: 10.1038/nmeth.3027. URL: http://dx.doi.org/10.1038/nmeth.3027.

[3] Thomas A Hopf et al. "Mutation effects predicted from sequence co-variation". In: *Nature Biotechnology* 35.2 (Jan. 2017), pp. 128–135. ISSN: 1546-1696. DOI: 10.1038/nbt.3769. URL: http://dx.doi.org/10.1038/nbt.3769.

[4] Joshua Meier et al. "Language models enable zero-shot prediction of the effects of mutations on protein function". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., 2021, pp. 29287–29303. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/f51338d736f95dd42427296047067694-Paper.pdf.

[5] Pascal Notin et al. "ProteinGym: Large-Scale Benchmarks for Protein Fitness Prediction and Design". In: *Advances in Neural Information Processing Systems 36*. NeurIPS 2023. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023, pp. 64331–64379. DOI: 10.52202/075280-2810. URL: http://dx.doi.org/10.52202/075280-2810.

[6] Nicholas C Wu et al. "Adaptation in protein fitness landscapes is facilitated by indirect paths". In: *eLife* 5 (July 2016). ISSN: 2050-084X. DOI: 10.7554/elife.16965. URL: http://dx.doi.org/10.7554/eLife.16965.

[7] Pascal Notin et al. "ProteinNPT: Improving Protein Property Prediction and Design with Non-Parametric Transformers". In: *Advances in Neural Information Processing Systems 36*. NeurIPS 2023. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023, pp. 33529–33563. DOI: 10.52202/075280-1458. URL: http://dx.doi.org/10.52202/075280-1458.

[8] Kaiyi Jiang et al. "Rapid in silico directed evolution by a protein language model with EVOLVEpro". In: *Science* 387.6732 (Jan. 2025). ISSN: 1095-9203. DOI: 10.1126/science.adr6006. URL: http://dx.doi.org/10.1126/science.adr6006.

[9] Vincent Q. Tran et al. "Rapid directed evolution guided by protein language models and epistatic interactions". In: *Science* 392.6798 (May 2026). ISSN: 1095-9203. DOI: 10.1126/science.aea1820. URL: http://dx.doi.org/10.1126/science.aea1820.

[10] Palash Sethi and Juannan Zhou. "An interpretable neural network unveils higher-order epistasis in large protein sequence-function relationships". In: (Sept. 2024). DOI: 10.1101/2024.09.22.614318. URL: http://dx.doi.org/10.1101/2024.09.22.614318.

[11] Steffanie Paul et al. "Combining Structure and Sequence for Superior Fitness Prediction". In: *GenBio Workshop at NeurIPS*. 2023.

[12] Lasse M. Blaabjerg et al. “SSEmb: A joint embedding of protein sequence and structure enables robust variant effect predictions”. In: *Nature Communications* 15.1 (Nov. 2024). ISSN: 2041-1723. DOI: 10.1038/s41467-024-53982-z. URL: http://dx.doi.org/10.1038/s41467-024-53982-z.

[13] Peter Groth et al. “Kermut: Composite kernel regression for protein variant effects”. In: *Advances in Neural Information Processing Systems 37*. NeurIPS 2024. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024, pp. 29514–29565. DOI: 10.52202/079017-0929. URL: http://dx.doi.org/10.52202/079017-0929.

[14] Yinghui Chen et al. “An end-to-end framework for the prediction of protein structure and fitness from single sequence”. In: *Nature Communications* 15.1 (Aug. 2024). ISSN: 2041-1723. DOI: 10.1038/s41467-024-51776-x. URL: http://dx.doi.org/10.1038/s41467-024-51776-x.

[15] John Jumper et al. “Highly accurate protein structure prediction with AlphaFold”. In: *Nature* 596.7873 (July 2021), pp. 583–589. ISSN: 1476-4687. DOI: 10.1038/s41586-021-03819-2. URL: http://dx.doi.org/10.1038/s41586-021-03819-2.

[16] Zeming Lin et al. “Evolutionary-scale prediction of atomic-level protein structure with a language model”. In: *Science* 379.6637 (Mar. 2023), pp. 1123–1130. ISSN: 1095-9203. DOI: 10.1126/science.ade2574. URL: http://dx.doi.org/10.1126/science.ade2574.

[17] Ruidong Wu et al. “High-resolution de novo structure prediction from primary sequence”. In: (July 2022). DOI: 10.1101/2022.07.21.500999. URL: http://dx.doi.org/10.1101/2022.07.21.500999.

[18] Jian Hu, Weizhe Wang, and Haipeng Gong. “Cerebra: a computationally efficient framework for accurate protein structure prediction”. In: (Feb. 2024). DOI: 10.1101/2024.02.02.578551. URL: http://dx.doi.org/10.1101/2024.02.02.578551.

[19] Salvatore Candido et al. “Language Modeling Materializes a World Model of Protein Biology”. In: *bioRxiv* (2026). DOI: 10.64898/2026.06.03.729735. URL: https://www.biorxiv.org/content/10.64898/2026.06.03.729735.

[20] Thomas Hayes et al. “Simulating 500 million years of evolution with a language model”. In: *Science* 387.6736 (Feb. 2025), pp. 850–858. ISSN: 1095-9203. DOI: 10.1126/science.ads0018. URL: http://dx.doi.org/10.1126/science.ads0018.

[21] Zhenda Xie et al. “mHC: Manifold-Constrained Hyper-Connections”. In: (2025). DOI: 10.48550/ARXIV.2512.24880. URL: https://arxiv.org/abs/2512.24880.

[22] Jürgen Haas et al. “Continuous Automated Model EvaluatiOn (CAMEO) complementing the critical assessment of structure prediction in CASP12”. In: *Proteins: Structure, Function, and Bioinformatics* 86.S1 (2018), pp. 387–398. DOI: 10.1002/prot.25431.

[23] Kadina E. Johnston et al. “A combinatorially complete epistatic fitness landscape in an enzyme active site”. In: *Proceedings of the National Academy of Sciences* 121.32 (July 2024). ISSN: 1091-6490. DOI: 10.1073/pnas.2400439121. URL: http://dx.doi.org/10.1073/pnas.2400439121.

[24] C. Anders Olson, Nicholas C. Wu, and Ren Sun. "A Comprehensive Biophysical Description of Pairwise Epistasis throughout an Entire Protein Domain". In: *Current Biology* 24.22 (Nov. 2014), pp. 2643–2651. ISSN: 0960-9822. DOI: 10.1016/j.cub.2014.09.072. URL: http://dx.doi.org/10.1016/j.cub.2014.09.072.

[25] Fabian Fuchs et al. "SE(3)-Transformers: 3D Roto-Translation Equivariant Attention Networks". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 1970–1981. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/15231a7ce4ba789d13b722cc5c955834-Paper.pdf.

[26] Yang Zhang and Jeffrey Skolnick. "Scoring function for automated assessment of protein structure template quality". In: *Proteins: Structure, Function, and Bioinformatics* 57.4 (2004), pp. 702–710. DOI: 10.1002/prot.20264. URL: https://doi.org/10.1002/prot.20264.

[27] Valerio Mariani et al. "lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests". In: *Bioinformatics* 29.21 (Aug. 2013), pp. 2722–2728. DOI: 10.1093/bioinformatics/btt473. URL: https://doi.org/10.1093/bioinformatics/btt473.

Acknowledgements

This work was supported by the National Natural Science Foundation of China grant 32671644 (H.G.) and 32500561 (Y.X.), the Ministry of Science and Technology of China grant 2023YFF1204400 (H.G.), the Beijing Frontier Research Center for Biological Structure, the Postdoctoral Fellowship Program of CPSF under Grant Number GZC20251844 (Y.X.), the China Postdoctoral Science Foundation under Grant Number 2026T190759 (Y.X.) and 2026M793139 (Y.X.) and the General Program of Hubei Provincial Natural Science Foundation of China grant 2026AFB647 (Y.X.).

Author contributions

H.G. and Y.X. conceived the study, proposed the methodology and developed the overall research framework. For Cerebra-Seq, H.G., W.W. and J.H. designed the initial framework, while W.W. implemented the model, performed model training and testing, and developed the web server. For Cerebra-Epistasis, Y.X. and Z.Y. designed the initial framework, while Z.Y. implemented the model and performed model training and testing. E.Y., Z.S. and S.Y. curated and preprocessed the datasets and conducted the benchmark evaluations.

H.G., Y.X., W.W., Z.Y. and Z.S. analyzed the results and wrote the manuscript. All authors reviewed and approved the final manuscript.

Competing interests

Yunxin Xu and Zimu Yu have submitted a patent application based on this Epistasis work. The remaining authors declare no competing interests.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- CerebraEpistasisSI.pdf