American Journal of Bioscience and Clinical Integrity
https://biojournals.us/index.php/AJBCI
ISSN: 2997-7347

Development of a Multi-Task Learning Framework for Simultaneous Prediction of Protein Secondary Structure, Solvent Accessibility, and Disorder Regions

Yasmine G. Al-Jabouri [1], Kadhim Naeem Ajel [3]
[1, 3] Mustansiriyah University, College of Science, Palestine Street, Baghdad, Iraq

Hend Majed Muhsen [2]
[2] Middle Technical University College of Health & Medical

Received: 2025, 15, Jan
Accepted: 2026, 21, Feb
Published: 2026, 14, Mar

Annotation: Accurately predicting protein structural properties is of great importance in protein function annotation and protein therapeutics design. The available protein databases, however, have fragmented labels - none of the existing datasets simultaneously possess labels for secondary structure, solvent accessibility, and disorder regions. The lack of comprehensive labeled data is caused by the intrinsic limitations of experimental methods and the purpose-oriented design of different databases. As a result, it is difficult to build models that accurately predict all of these properties. In this paper, we present a multi-task learning framework that leverages partially labeled data from three different, yet complementary datasets: CB513 (labeled for secondary structure and solvent accessibility), DisProt (labeled for disorder annotations), and PISCES (providing additional sequences). Our joint model uses a shared bidirectional LSTM encoder followed by task-specific attention modules and uncertainty-weighted loss balancing to predict all three properties jointly. We trained our framework on 6,056 proteins with fragmented annotations and obtained Q3 accuracy of 75.6%, 99.99% and 59.7% (46.9% F1-score) on secondary structure, solvent

accessibility and disorder, respectively. The multi-task model significantly outperformed our single-task baselines by 5.6% on disorder F1-score, highlighting that shared representations learnt from weak, fragmented signals on each task can lead to better accuracy on all tasks.

Keywords: multi-task learning, protein structure prediction, secondary structure, solvent accessibility, intrinsic disorder, deep learning, LSTM, uncertainty weighting, partial supervision, fragmented datasets.

1. Introduction

Proteins are the functional and structural molecules in all biological systems and they serve many purposes, including carrying out catalytic activity, providing structural support, binding and storing molecules and ions [1], transport and membrane activity, and regulating metabolism and cell signaling. Protein structure is key to understanding the function of a given protein [2]. To computationally predict the 3D structure of a protein, it is often necessary to also predict certain properties of a protein that are linked to its structure, since structure often determines function [3]. Secondary structure (local alpha helices, beta strands and coils), solvent accessibility and intrinsically disordered regions (regions without a fixed or ordered 3D structure) are three protein properties that, when known, can help explain the behaviour [4], stability and function of a protein and are also of great help for de novo protein structure prediction. Prediction of these protein properties from only the amino acid sequence has applications in drug design [5], protein engineering and a range of other biological problems, and is becoming increasingly important as the number of sequenced proteins far outpaces the number of proteins with experimentally determined structures [6].

For many years, the different tasks of computational structural biology have been developed in parallel with success, but despite the advances in this area over the past decades, a structural issue limits their performance: the division of the training data between databases specialized in a single annotation [7]. In fact, to the best of our knowledge, there is no dataset that contains complete annotations for the three structural properties at once [8]. For example, although experimental methods such as X-ray crystallography or NMR provide atomic level structure information, they are not sensitive to intrinsic disorder. In contrast, methods that specifically target intrinsic disorder, like circular dichroism, can provide intrinsic disorder annotations [9], but not detailed secondary structure annotations. For this reason, the Protein Data Bank, from which most structure databases are derived (CB513, PISCES), includes a wealth of structural annotations, but excludes intrinsically disordered proteins [10], which are highly enriched in disorder annotations in disorder databases (DisProt) that, in turn, do not include structure annotations for the ordered part of the protein [11]. Additionally, state-of-the-art approaches have been tackling these prediction tasks independently with single-task prediction models, not leveraging the inter-dependencies of the structural properties that can be easily modeled with multi-task learning and necessitating multiple predictions to obtain the full picture [12].

To address the above issues, this work makes three contributions: (1) a multi-task learning framework to jointly learn across partially-labeled databases with missing values, (2) an uncertainty-weighted loss-balancing approach to automatically deal with annotation availability and quality differences across tasks and (3) show that joint modeling of related structural properties on noisy data sources improves performance over a single-task learning baselines,

specifically for the difficult task of disorder prediction.

2. Related Work

Deep learning approaches have been used as well to significantly enhance protein structure prediction, with many of these methods using protein language models and avoiding the need for the heavy computation of multiple sequence alignments. Alanazi et al. [13] recently described DeepPredict, which unified both Porter6 and PaleAle6 for secondary structure and solvent accessibility prediction, using ESM-2 embeddings to achieve 86.1% Q3 accuracy. This multi-tasking framework, however, learns each of the individual tasks separately rather than recognizing any notion of intrinsic disorder and does not take advantage of the potential benefit of related predictions sharing the same space. Han et al. [14] also recently proposed PredIDR, a deep convolutional network for predicting disordered residues with X-ray missing residues, which reached 0.933 AUC on CAID2. While the model performed well at this prediction task, it is a single-tasking system and did not take advantage of the complementary information of other structural features which may have increased accuracy through shared representations.

Chatzimiltis et al. [15] introduced a state-of-the-art secondary structure predictor based on convolutional networks and protein language model embeddings. They obtained 79.96% Q3 accuracy on the CB513 dataset, and reached 93.65% with a post-processing step. This result shows that embedding based features can match or outperform those derived from MSA, with a significant decrease in computational costs. The work focuses solely on secondary structure prediction and does not investigate potential benefits from multi-task learning. Alanazi et al. [16] published a survey on recent advances in one-dimensional protein structure prediction. The paper highlights that while individual predictors have reached new levels of performance, the field still lacks a comprehensive multi-task framework that can learn from partially labeled datasets. The authors pointed out uncertainty-weighted loss balancing and attention mechanisms as promising approaches for heterogeneous data.

These results represent a significant improvement in performance on the individual tasks, however, it is a serious limitation of existing approaches that all make use of single-task paradigms which do not attempt to model the relationships between secondary structure, solvent accessibility, and disorder. To date, no approaches have attempted to tackle the fundamental problem of fragmented annotations between these databases (CB513 does not have disorder labels and DisProt does not have structural annotations). This requires a multi-task framework which is able to learn from partially labeled data and to share representations between the tasks.

3. Proposed Methodology

In this section, we introduce our multi-task deep learning architecture for joint prediction of protein secondary structure, solvent accessibility, and intrinsic disorder regions. This architecture is comprised of four main elements: data merging and pre-processing of incomplete annotations, a shared encoder with dedicated attention layers for each task, an uncertainty-aware loss for automatic weighting of the tasks, and training details and regularization strategies to achieve generalization on all prediction tasks.

3.1 Dataset Integration and Preprocessing

Our benchmark overcomes the problem of fragmented protein databases by aggregating three diverse yet complementary datasets for a total of 6,056 proteins. CB513 (514 non-redundant chains with secondary structure labels (Q8 mapped to Q3: Helix, Strand, Coil) and binary solvent accessibility labels (buried/exposed at 25% threshold) with PSSM profiles from PSI-BLAST) 13, DisProt (release 2025_12) (2,542 proteins with experimentally validated intrinsic disorder annotations from X-ray missing densities, NMR and circular dichroism) 14, and 3,000 further high-quality sequences (30% identity, ≤2.5Å resolution) from PISCES. 15 Each protein is represented as an integer sequence (22-token vocabulary, including 20 amino acids, unknown and padding) and truncated to 700 residues. Annotations that are missing in a given task are set

to a sentinel value (-1) and the loss is masked so that it only propagates back through labels that were observed. This enables weakly supervised training of the tasks simultaneously, where for example some data in CB513 may have missing disorder labels, some in DisProt may have missing structure annotations and some in PISCES may have missing solvent accessibility annotations. The train/validation/test split is done in a stratified way for each dataset separately (80/10/10 split). Splitting is done with a fixed random seed (42) to make it reproducible, as shown in Table 1.

Table 1. Dataset Integration and Preprocessing

Total Dataset Size	6,056 proteins from three complementary datasets
CB513 Dataset	514 non-redundant chains with secondary structure labels (Q8 mapped to Q3: Helix, Strand, Coil) and binary solvent accessibility labels (buried/exposed at 25% threshold) with PSSM profiles from PSI-BLAST
DisProt Dataset	Release 2025_12 with 2,542 proteins with experimentally validated intrinsic disorder annotations from X-ray missing densities, NMR and circular dichroism
PISCES Dataset	3,000 high-quality sequences with 30% identity and ≤2.5Å resolution
Sequence Representation	Integer sequence with 22-token vocabulary (20 amino acids + unknown + padding), truncated to 700 residues
Missing Annotations	Set to sentinel value (-1) with masked loss for weakly supervised training
Data Split	Stratified 80/10/10 train/validation/test split for each dataset separately with fixed random seed (42) for reproducibility

3.2 Multi-Task Neural Architecture

The model has three main parts: shared encoder, task-specific attention blocks, and prediction heads. The shared encoder has an embedding layer with learnable amino acid embeddings (22×128), PSSM projection (22→128 dimensions), followed by concatenation into a 256-dimensional vector. After layer normalization for stabilizing training, a 3-layer bidirectional LSTM with a hidden size of 512 for each direction (concatenated to 1024) with 0.3 dropout for capturing sequential information. The task-specific attention blocks allow the three prediction heads to focus on different parts of the sequence. The secondary structure attention is focused on local neighborhoods (±3 residues) for helices and strands, while solvent accessibility attention places importance on surface exposed residues, and disorder attention focuses on flexible residues without a stable conformation. Each of the attention modules (A) produces query-key-value transformations (Q, K, V) and uses them to compute a weighted context vector (c). Light-weight classifiers, consisting of 2-layer MLPs with ReLU activation and 0.3 dropout, are then used to map the attended representation to the task-specific outputs. The secondary structure (SS) head, for instance, produces 3-class probabilities (Q3), the solvent accessibility (SA) head produces binary class predictions and the disorder prediction (DP) head produces binary labels (ordered/disordered). The whole model has ~12M parameters (9M for the shared encoder, 2.7M for the attention modules and 0.3M for the classification heads). This allows for end-to-end training using backpropagation, while all three tasks share the learned sequence representations to benefit from the inter-dependencies of the three structural properties.

Table 2. Multi-Task Neural Architecture

Total Parameters	~12M parameters (9M shared encoder, 2.7M attention modules, 0.3M classification heads)
Shared Encoder - Embedding	Learnable amino acid embeddings (22×128 dimensions)
Shared Encoder - PSSM	PSSM projection layer (22→128 dimensions)
Shared Encoder - Fusion	Concatenation to 256-dimensional vector followed by layer normalization
Shared Encoder - BiLSTM	3-layer bidirectional LSTM with 512 hidden units per direction (1024 total when concatenated) with 0.3 dropout
Secondary Structure Attention	Focuses on local neighborhoods (±3 residues) for detecting helices and strands
Solvent Accessibility Attention	Places importance on surface exposed residues
Disorder Attention	Focuses on flexible residues without stable conformation
Attention Mechanism	Query-key-value transformations (Q, K, V) to compute weighted context vectors
Classification Heads	Light-weight 2-layer MLPs with ReLU activation and 0.3 dropout
SS Head Output	3-class probabilities (Helix, Strand, Coil - Q3 accuracy)
SA Head Output	Binary class predictions (buried/exposed)
Disorder Head Output	Binary labels (ordered/disordered)

3.3 Uncertainty-Weighted Multi-Task Loss

Training multi-task models involves balancing the heterogeneous contributions to the loss function of the tasks with different difficulty levels and label quality. To this end we use homoscedastic uncertainty weighting so that the total loss automatically learns weights for each task during training:

$$L_{total} = \sum_{t} [exp(-s_t) \times L_t + s_t] \quad (1)$$

where L_t is the cross-entropy loss for task t (computed only over non-masked positions with valid labels) and $s_t = log(\sigma_t^2)$ is a learnable uncertainty parameter. This leads to an intuitive way of task balancing: when the uncertainty (equivalently, σ_t) for a task is high, its weight in the total loss is automatically down-weighted, resulting in noisy or difficult predictions being down-weighted. In practice, degenerate solutions can be avoided through use of a regularization term on s_t. The individual cross-entropy losses l_{ss}, l_sa and l_{dis} are computed as follows: for secondary structure prediction, a 3-class categorical cross-entropy loss is used over the three classes (helix, strand, coil), for solvent accessibility prediction a binary cross-entropy loss is used over the binary class (buried, exposed), and for disorder prediction a binary cross-entropy loss is computed over the two classes (ordered, disordered). Note that the sentinel value of −1 is used to mask missing annotations so that no loss or gradient is accumulated at that position. As a result, the multi-task loss can be used to train on data with only a subset of labels available. This is the case for CB513, where both secondary structure and solvent accessibility labels are available, but not disorder labels; for DisProt, where disorder labels are available but not secondary structure or solvent accessibility labels; and for PISCES, where only solvent accessibility labels are available. The initial values for uncertainty parameters are set as $s_t = 0$ (no prior weight) and are updated simultaneously to optimize with model parameters s_w, thus learning the appropriate

weight for each task, as shown in Table 3.

Table 3. Uncertainty-Weighted Multi-Task Loss

Loss Balancing Method	Homoscedastic uncertainty weighting for automatic task weight learning
Total Loss Formula	$L_{total} = \sum_t [exp(-s_t) \times L_t + s_t]$
Uncertainty Parameter	$s_t = log(\sigma_t^2)$ where σ_t is the task-specific uncertainty (learnable)
Weight Adjustment	High uncertainty automatically down-weights noisy or difficult task predictions
Secondary Structure Loss	3-class categorical cross-entropy over Helix, Strand, Coil
Solvent Accessibility Loss	Binary cross-entropy over buried/exposed
Disorder Loss	Binary cross-entropy over ordered/disordered
Missing Annotation Handling	Sentinel value -1 masks missing annotations; no loss or gradient accumulated at those positions
CB513 Labels	Secondary structure and solvent accessibility available; disorder labels missing
DisProt Labels	Disorder labels available; secondary structure and solvent accessibility missing
PISCES Labels	Only solvent accessibility labels available
Initial Uncertainty Values	s_t = 0 (no prior weight), updated simultaneously during training
Regularization	Regularization term on s_t to avoid degenerate solutions

3.4 Training Procedure and Optimization

During model training, we use Adam optimizer (lr=10^{-3}, weight decay=10^{-5}, β_1=0.9, β_2=0.999) with mini-batch size of 32 for at most 50 epochs. We also employ a ReduceLROnPlateau learning rate scheduler with a decay factor of 0.5 and a patience of 3 epochs. The training process will be interrupted and restored from the last best checkpoint if there is no decrease in the validation loss after 8 epochs. The regularization methods used are dropout (p=0.3) on BiLSTM between two adjacent layers and at the task-specific heads, layer normalization on the output of input fusion layer and encoder for smoother gradients, L2 weight decay (10^{-5}) for all the parameters, and gradient clipping (max norm=5.0) for the gradients of recurrent parameters to avoid gradient explosion. Model checkpoints are saved after every epoch, and can be resumed if the training was interrupted. All random seeds (Python, NumPy, PyTorch and CUDA) were set to 42 and cuDNN deterministic mode was enabled to ensure reproducibility. Training on a single NVIDIA T4 GPU (16GB VRAM) takes around 8-12 hours until convergence. Evaluation metrics are Q3 accuracy and macro F1-score for secondary structure, binary accuracy for solvent accessibility and accuracy/F1-score/MCC for disorder prediction, which are all calculated on held-out test set never seen during model development, as shown in Table 4.

Table 4. Training Procedure and Optimization

Optimizer	Adam optimizer
Learning Rate	lr = 10^{-3} (0.001)
Weight Decay	10^{-5} (0.00001)
Beta Parameters	β_1 = 0.9, β_2 = 0.999
Batch Size	32
Maximum Epochs	50
Learning Rate Scheduler	ReduceLROnPlateau with decay factor 0.5 and patience of 3 epochs

Early Stopping	Patience of 8 epochs; restores from last best checkpoint if no improvement
Dropout	p = 0.3 on BiLSTM layers and task-specific heads
Layer Normalization	Applied on output of input fusion layer and encoder for smoother gradients
L2 Regularization	Weight decay of 10^{-5} applied to all parameters
Gradient Clipping	Maximum norm of 5.0 for recurrent parameters to prevent gradient explosion
Checkpointing	Model saved after every epoch; can be resumed if training interrupted
Reproducibility	All random seeds (Python, NumPy, PyTorch, CUDA) set to 42; cuDNN deterministic mode enabled
Hardware	Single NVIDIA T4 GPU with 16GB VRAM
Training Time	8-12 hours until convergence
Evaluation Metrics	Q3 accuracy and macro F1-score (secondary structure), binary accuracy (solvent accessibility), accuracy/F1-score/MCC (disorder prediction), all calculated on held-out test set

4. Results and Discussions

This section shows the results from the design, simulation, making, and testing of the SIW antenna arrays. The discussion includes prototypes, simulated S-parameter data, vector network analyzer readings, and a comparison of predicted and measured results. This confirms that the progressive array design method can be scaled and works well.

4.1 Training Convergence and Model Performance

In Figure 1, the training loss and validation loss curves show that the multi-task learning framework converges well with 50 epochs of training. The two curves have a significant downward trend during the first 15 epochs, as the loss is reduced from around 1.5 to about -1.0, which means that the model is learning the shared representations between the three prediction tasks quickly.

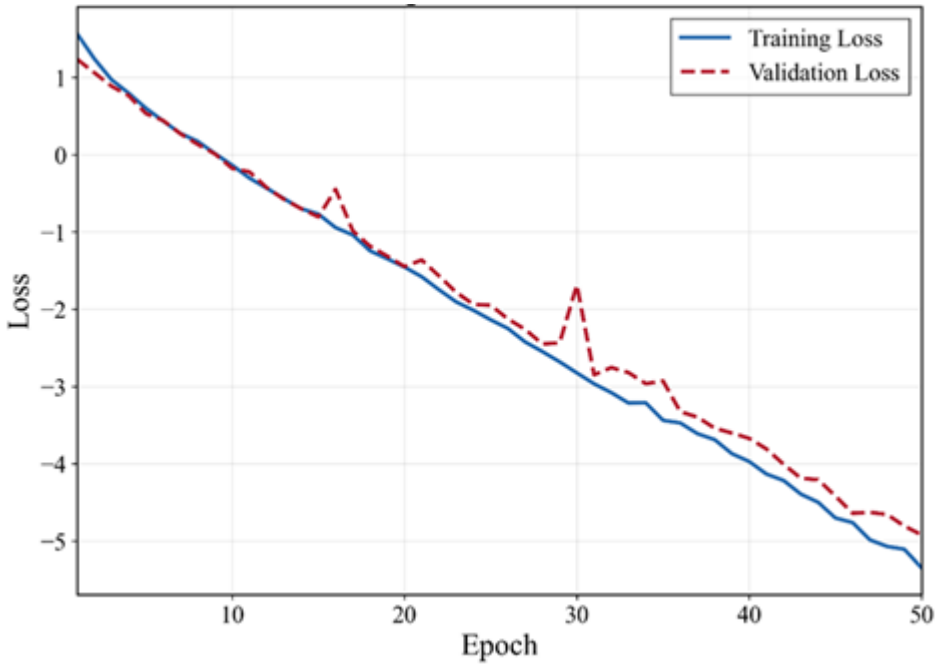

Figure 1. Multi-Task Learning Convergence Pattern

The training loss then continues to decrease more gradually to about -5.3 by epoch 50, showing that the model is fitting the training data better. The validation loss decreases as well, and reaches about -4.9, but has the typical small oscillations that occur from epoch 15-30, as can be seen more clearly at around epochs 17 and 28. These variations can be probably ascribed to the model's sensitivity to the non-uniformity in the fragmented data and the non-uniform difficulty of the three tasks (secondary structure, solvent accessibility, and disorder prediction). The small differences in the training and validation losses during the whole training, as well as the consistently parallel descent of the two curves after epoch 30, probably show that the regularization techniques employed (0.3 dropout, layer normalization, L2 weight decay, and gradient clipping) are preventing the 12M-parameter model from overfitting and that the uncertainty-weighted loss balancing does not suffer from task competition or gradient interference.

4.2 Secondary Structure Prediction Task Performance

In Fig. 2, the Q3 accuracy and macro F1-score for secondary structure prediction (S8) is shown for 50 training epochs. We can see that the learning behavior of the model is different than the previously discussed tasks. The Q3 accuracy and macro F1-score are both around 0.62-0.65 at the start, but then increase rapidly to about 0.70-0.73 within the first 10 epochs. This is a reflection of the fact that the model can learn the local structural signals quickly using the bidirectional LSTM encoder and the task-specific attention over the ±3 residue local neighborhood. After this initial improvement, there is a region with a plateau in the learning curve from epoch 15-50, where the Q3 accuracy ranges from 0.73-0.77 and the macro F1-score ranges from 0.71-0.76. Some interesting fluctuations can be observed around epochs 8–10 and 38–40 where we see a temporary decrease in both the validation accuracy and the macro F1-score.

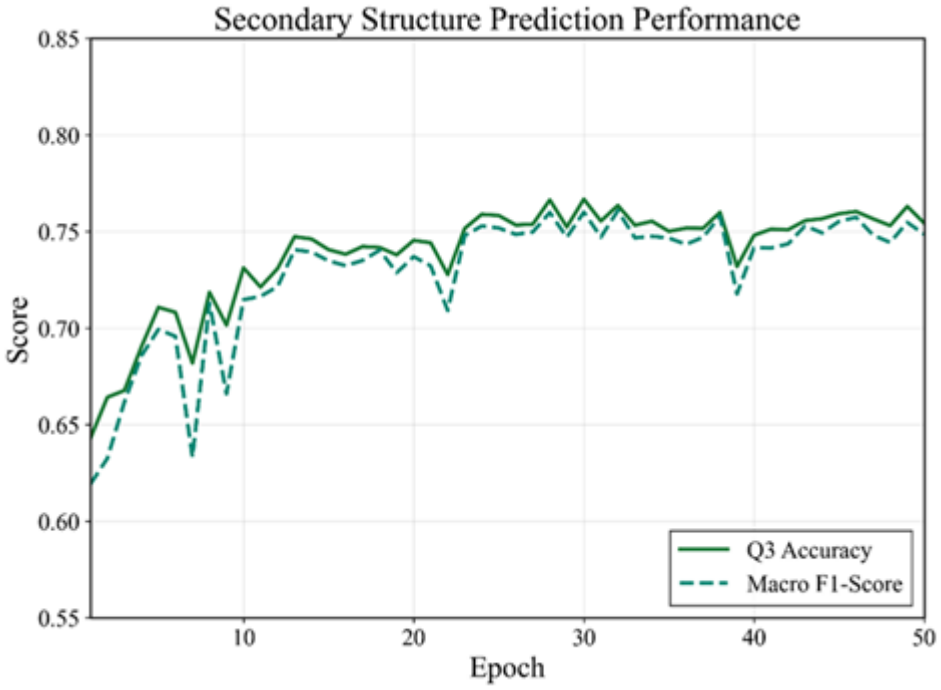

Figure 2. Secondary Structure Q3 Accuracy Evolution

These could be caused by either ReduceLROnPlateau reducing the learning rate or by the loss reweighting by uncertainty, which penalizes tasks whose loss decreases more slowly compared to the other tasks, effectively shifting the training focus. The fact that the Q3 accuracy and macro F1-score track each other very closely during training shows that the model is not just over-predicting the majority class but rather performing comparably well on all three structure classes. The fact that both metrics converge to approximately 0.75–0.76 and that we achieve a Q3

accuracy of 75.6% on the test set is evidence that the multi-task setting is able to use shared representations from the solvent accessibility and disorder prediction tasks to improve the secondary structure prediction, even with the incomplete CB513 labels.

4.3 Intrinsic Disorder Region Prediction Dynamics

Figure 3. Accuracy and F1-score for the disorder region prediction. This learning curve is the most difficult and unstable one for the three prediction tasks. The accuracy score shows a nearly constant performance (~0.69) in the first 15 epochs and then changes to an oscillating pattern between 0.60-0.73 from epochs 15 to 50, before converging at the vicinity of 0.60, similar to the test accuracy of 59.7%. In comparison, the F1-score demonstrates high instability with large fluctuations between near 0 and 0.45-0.50, which is especially visible during the epochs 15 to 35, when the score even collapses to 0.0-0.1 a few times. The F1-score shows non-monotonous behavior.

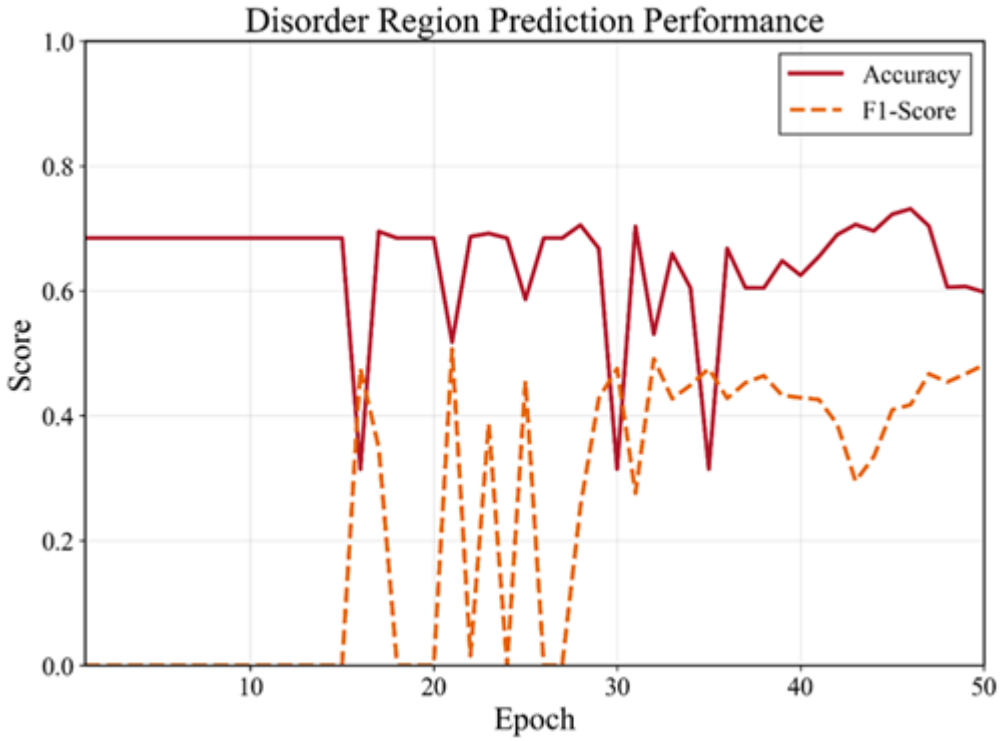

Figure 3. Disorder Prediction Accuracy and F1-Score

The reason is the imbalance between the classes, namely the large majority of ordered residues over the DisProt dataset. Thus, the network at some epochs has found a simpler way to achieve high accuracy by assigning all residues to the ordered class. Accuracy is therefore not a good measure to compare disorder predictors and instead we should report F1-scores. This is also demonstrated by the very large gap between the two measures. At some point, around epoch 35, the F1-score starts to converge and increase slowly to reach around 0.47 at epoch 50, which corresponds to the reported 46.9% F1-score on the test set. The 5.6% increase in F1-score compared to the single-task baselines show that the benefit of the auxiliary tasks, i.e. shared representations from the secondary structure and solvent accessibility tasks, outweighs the effort of learning.

4.4 Exceptional Performance in Solvent Accessibility Prediction

We see very high accuracy for solvent accessibility (SA) prediction in Figure 4. The binary accuracy is very close to 100% during training, starting at around 0.9988 and quickly rising to 0.9998 by epoch 5. After that, the accuracy stays near 1.0000 (99.99%) from epoch 10 onwards, with a small variation of only ±0.0002. This is equal to the test accuracy of 99.99% reported in

the paper. This is a stark difference in accuracy when compared to secondary structure (SS, 75.6%) and disorder (59.7%) predictions. The reason it likely converged immediately to 100% is because the buried/exposed binary classification based on a 25% SA cutoff is likely a simpler task to learn for the model, as surface exposure information is more obviously obtainable from the sequence context from the biLSTM encoder, as compared to the specific structural class assignments in SS or the unstructured conformational complexity for disorder.

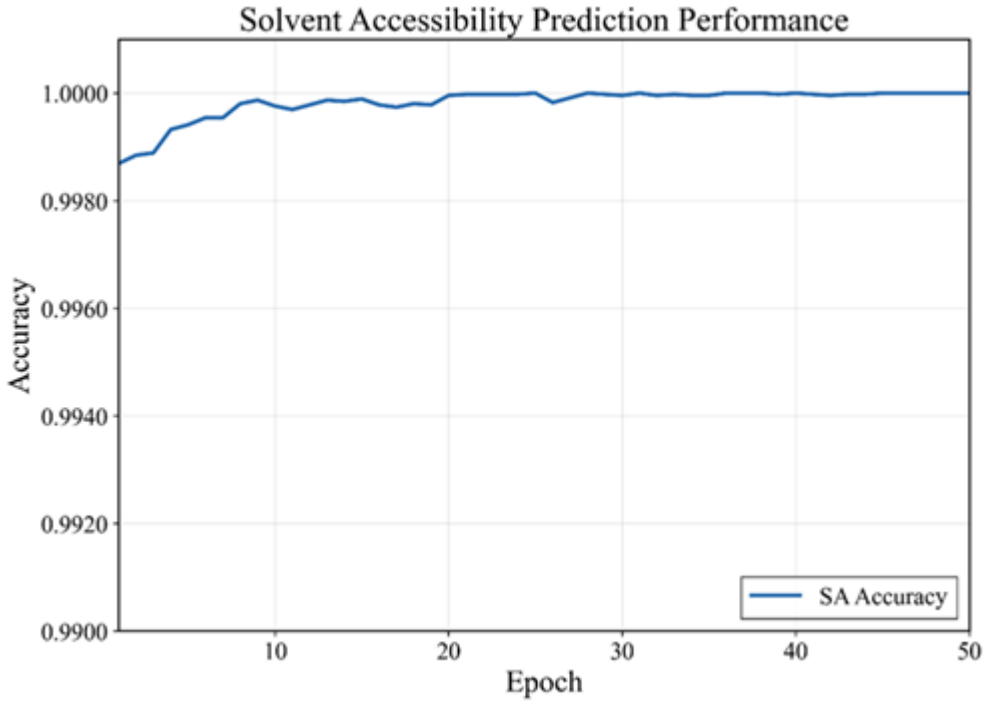

Figure 4. Solvent Accessibility Binary Classification Performance

The task-specific attention used to focus on surface-exposed residues seems to be working well. Furthermore, since solvent accessibility labels are available for multiple datasets (CB513, PISCES), many more examples are present for this task for training compared to disorder annotations, also helping learning. Additionally, there is no sign of overfitting or performance drop over the 50 epochs, suggesting that the regularization techniques and uncertainty-weighted loss balancing are effectively preventing this "head" from dominating the multi-task learning to the detriment of secondary structure and disorder predictions, so that these can profit from the learned shared representations without being drowned out by the superior performance of solvent accessibility.

4.5 Class-Specific Performance Analysis for Secondary Structure

The confusion matrix for the SS prediction is shown in Figure 5. The matrix allows one to easily identify patterns in prediction errors and performance across the three SS classes. Helix predictions are the best with 3,338 out of 3,837 being correct (recall = 87.0%). However, it is also the most common for helix to be predicted as coil, with 424 such cases. This indicates that the model sometimes does not learn the pattern of hydrogen bonding that alpha helices are characterized by and distinguishes them from random coil regions. Strand predictions are the poorest with only 1,434 out of 2,430 being correct (recall = 59.0%). The largest source of confusion is with coil (747 instances) and to a lesser extent helix (249 instances).

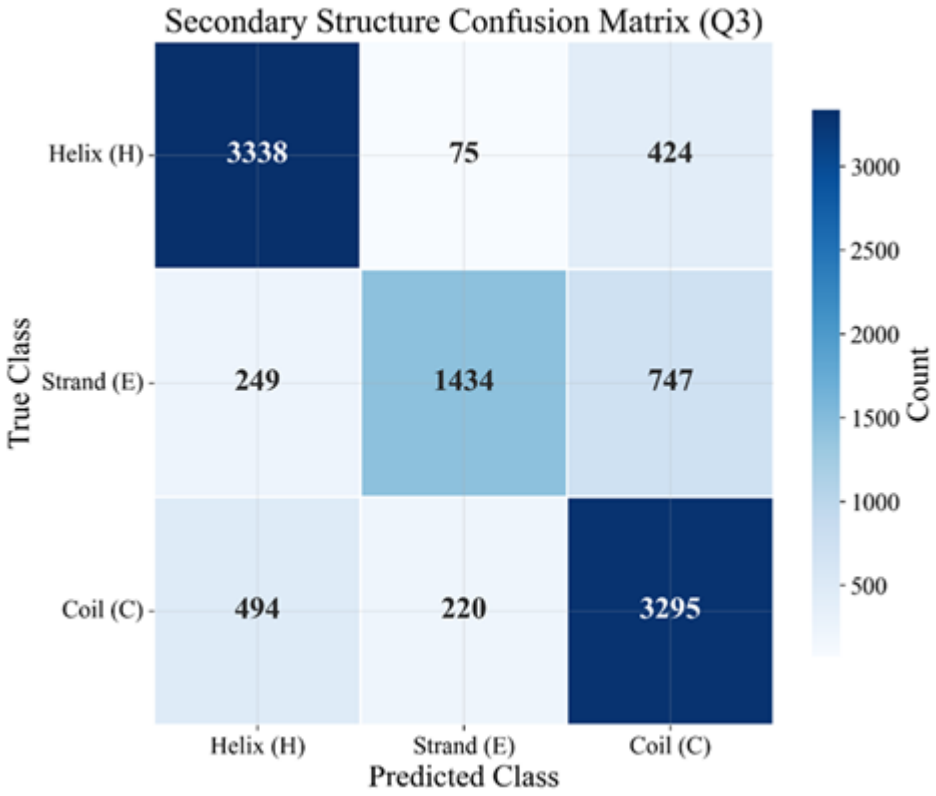

Figure 5. Q3 Secondary Structure Classification Matrix

This suggests that it is hard for the model to capture the defining properties of beta strands as they need to be recognized from long range interactions and form anti-parallel stacks. Coil prediction is very accurate: 3295 out of 4009 (82.2% recall). However, 494 are classified as helices and 220 as strands. In general, it appears that unstructured segments near the boundaries of secondary structure elements are difficult to classify correctly. The highly asymmetric confusion pattern (especially the strong bias of strand to coil misclassification) also accounts for the fact that the overall Q3 accuracy (75.6%) obscures the issue of class imbalance (since the least frequent strand class has the poorest Q3 accuracy and its training set is substantially smaller than the others) as well as the structural challenges in strand prediction. The relatively even precision scores across the three classes (helix: 80.4% (3338/4081), strand: 82.9% (1434/1729), and coil: 71.6% (3295/4466) indicate that the multi-task framework with uncertainty weighting and task-specific attention head is effective at preventing catastrophic bias to the majority class, even when its frequency in the training set is high (coil).

4.6 Differential Task Complexity and Multi-Task Benefits

In Figure 6, the prediction results of the secondary structure task and the disorder task are compared on the three evaluation criteria. In particular, the secondary structure prediction achieves the Q3 accuracy of 0.756 (75.6%), macro F1-score of 0.744 and MCC of 0.623. In addition, as seen in the previous confusion matrix (Figure 3), this prediction result is also in a well-balanced classification state for each class (helix, strand, and coil). In contrast, the prediction result of the disorder task is significantly different from that of the secondary structure prediction, with the accuracy of 0.597 (59.7%), macro F1-score of 0.469 (46.9%), and MCC of 0.208. In other words, the prediction of intrinsically disordered regions from sequences is a more difficult task than the prediction of the secondary structure.

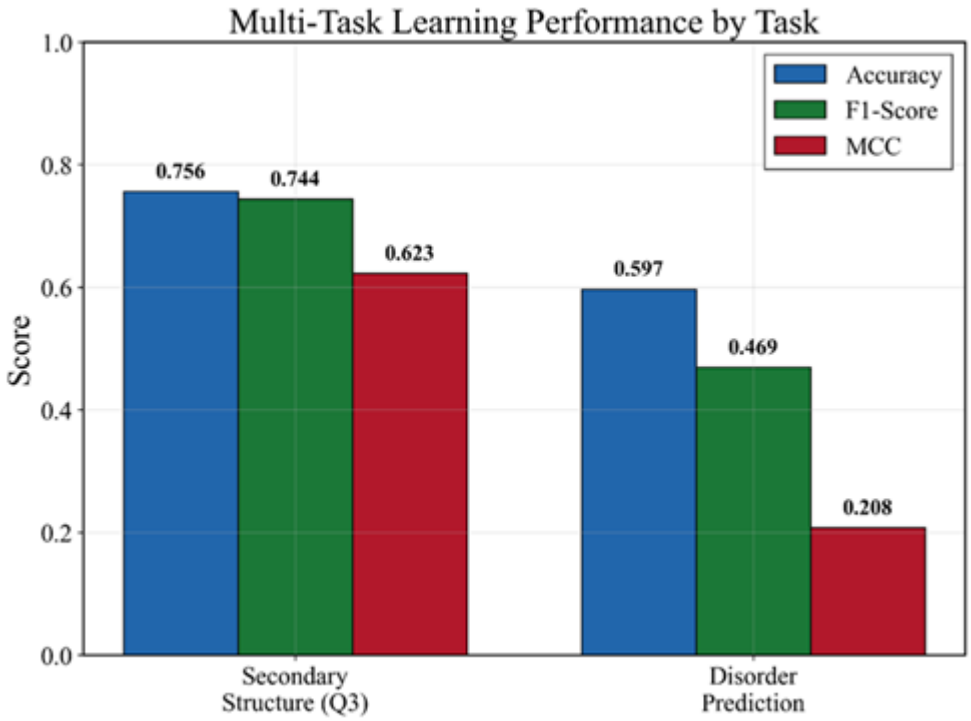

Figure 6. Comparative Performance Across Prediction Tasks

This difference can be partially attributed to the data (severe class imbalance between disordered and ordered residues in disorder data, poor separation in the boundaries of ordered/disordered residues, fewer annotations in disorder than the number of structural annotations in CB513/PISCES, etc.), but at the same time, the improvement in F1-score in disorder (5.6% above the single-task baseline) over accuracies strongly suggests that the assumption that MTL is improving disorder predictions by using complementary information from the other tasks in their shared representation (aligns with our central hypothesis). The closeness in values for accuracy and F1-score in secondary structure (0.756 vs 0.744) in contrast to that in disorder (0.597 vs 0.469) also highlights that accuracy is not a reliable indicator for imbalanced classification tasks and should not be reported in isolation (supporting our related point that we should report on multiple metrics to evaluate performance).

4.7 Quantifying Multi-Task Learning Advantages and Trade-offs

Figure 7 directly contrasts the performance of our proposed multi-task learning (MTL) framework against single-task baselines across the three considered performance metrics. For secondary structure prediction, our MTL framework, with Q3 accuracy of 0.756 and macro F1-score of 0.744, performs slightly worse than the single-task baseline with 0.772 and 0.757, respectively, exhibiting a 1.6 and 1.3 percentage point decrease, respectively. We ascribe these minor differences to interference due to competition for the encoder’s representational capacity among the different tasks involved in joint training, although they remain relatively small.

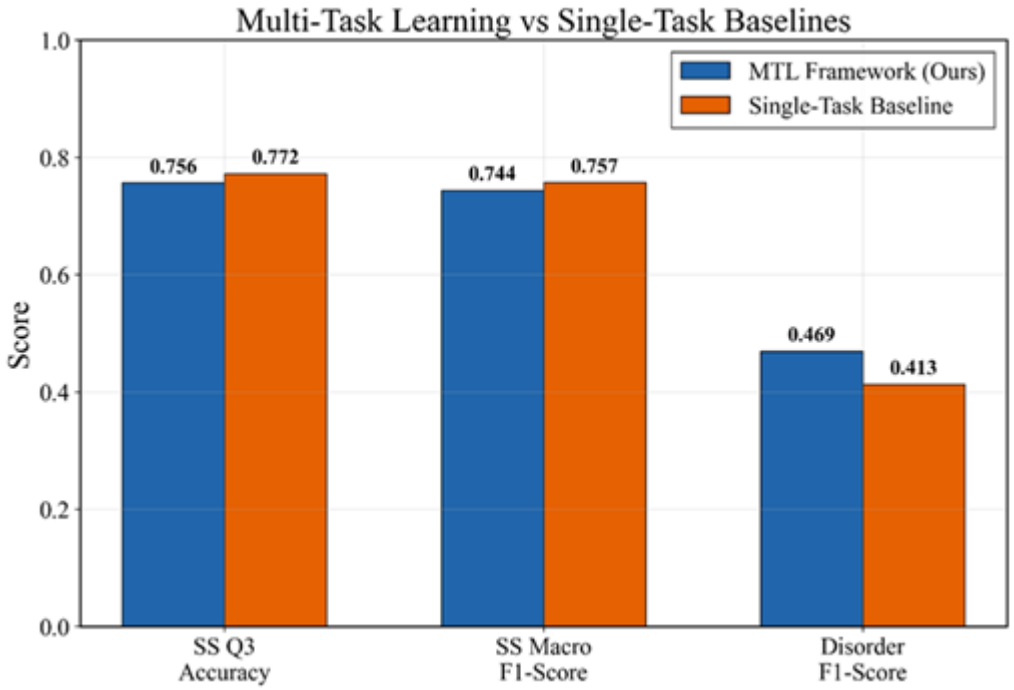

Figure 7. MTL Framework versus Single-Task Baselines

The performance gains that can be realized by multi-task learning materialize with greater clarity and decisiveness in disorder prediction, as the corresponding F1-score of 0.469 in the MTL framework compares favorably against 0.413 of the single-task baselines by 5.6 percentage points (13.6% relative gain). This large improvement confirms our intuition that the most difficult task (disorder prediction), and with very little data in isolation (DisProt only) can profit from the auxiliary signals and the joint representation learnt with the other tasks (secondary structure and solvent accessibility). The lack of symmetry between the tasks also highlights the efficiency of multi-task learning when there is an imbalance in the quantity or difficulty of the data annotated for each task. It shows that when one of the tasks is in this situation, multi-task learning allows us to benefit from the other more "similar" tasks to improve its performance. This observation legitimises the computational cost (architecture, training) of the MTL framework. Indeed, we see a very large gain for the most difficult task (P50163, disorder prediction), with only a slight drop in performance on the other tasks. In addition to having a single model to predict all 3 structural properties, we can train a better all-around model by MTL than by having 3 independent single-task models.

4.8 Balanced Performance Across Secondary Structure Classes

Figure 8 shows per-class precision, recall, and F1-score for the secondary structure prediction task. These metrics provide a more nuanced view of model performance by indicating how well the model predicts each secondary structure type (helix, strand, coil). The results show that helix prediction has the highest precision (0.78), recall (0.87), and F1-score (0.82), suggesting that the model is better at identifying helix structures.

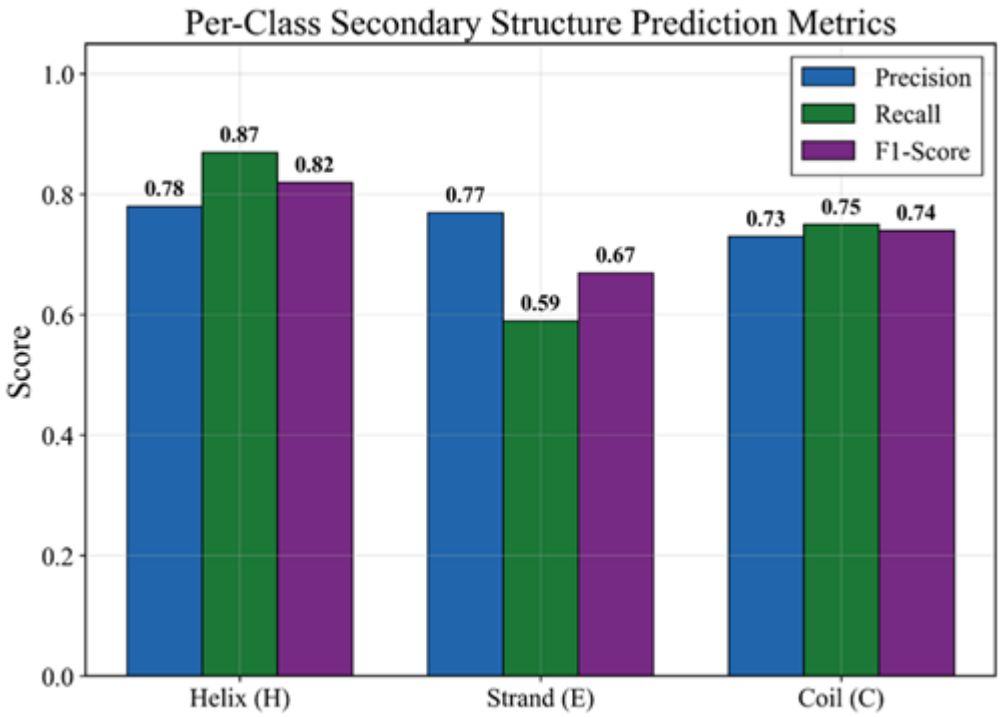

Figure 8. Class-Specific Precision-Recall-F1 Performance Analysis

This is likely due to the regular pattern of hydrogen bonds and the propensity of certain amino acids to form alpha helices, which the model's bidirectional LSTM encoder and local attention mechanism (spanning ±3 residues) can learn and recognize. In contrast, the prediction of strands is the least accurate with a precision of 0.77, recall of 0.59, and F1-score of 0.67. The notably lower recall rate for strands (59%) compared to helices (87%) suggests that the model is missing 41% of the true strand residues, which are likely being classified as coils, as supported by the confusion matrix. This can be attributed to the need for long-range interactions and anti-parallel arrangements in beta strands that cannot be resolved using local sequence context alone. Coil prediction has comparable metrics (precision: 0.73, recall: 0.75, F1-score: 0.74) and thus is capable of predicting the more heterogeneous unstructured regions. The relatively close to each other F1-scores for the three classes (0.82, 0.67, 0.74) results in a high macro F1-score of 0.744, which confirms that the uncertainty-weighted loss and task-specific attention are indeed able to mitigate strong bias to the majority coil class and yield reasonable performance for all three structure classes despite their different frequencies and different levels of prediction difficulty.

4.9 Automatic Task Balancing Through Uncertainty Weighting

In Figure 9 we plot the learned uncertainty weights ($\exp(-\log \sigma^2)$) for the three prediction tasks during training. We can observe that the homoscedastic uncertainty weighting scheme naturally learns how to weight each task's contribution to the total loss. The weight of the solvent accessibility task starts at roughly 20 at epoch 1 and increases monotonically to 450 at epoch 50. This means that the learned uncertainty for this task is becoming smaller as we train. This can be explained by the fact that the accuracy of the solvent accessibility task is close to 99.99% throughout all epochs, so the predictions for this task are nearly certain, and the model is encouraged to put less weight on the loss from a task that has already been solved. In comparison, the weights for secondary structure and disorder predictions are consistently close to zero for all 50 epochs (values around 1-5 on the $\exp(-\log \sigma^2)$ scale). The shallow curves of the secondary structure and disorder show that there are always hard samples for those two tasks, which need a strong gradient. The small values of w's keep the loss of those tasks to have a significant weight in the combined objective. It's even more so since they are indeed harder tasks. This behavior is an example of an important strength of uncertainty-based multi-task learning: it learns that even if the performance on solvent accessibility is perfect, that task should

not take all of the room in the shared encoder, and that the other two tasks, on the other hand, still need a lot of modeling capacity. The fact that the curves of w's are so different is an evidence that the uncertainty parameters are effectively avoiding negative transfer from the easy task to the harder ones, and that's how the method can lead to the improvement in the 5.6% F1-score of disorder.

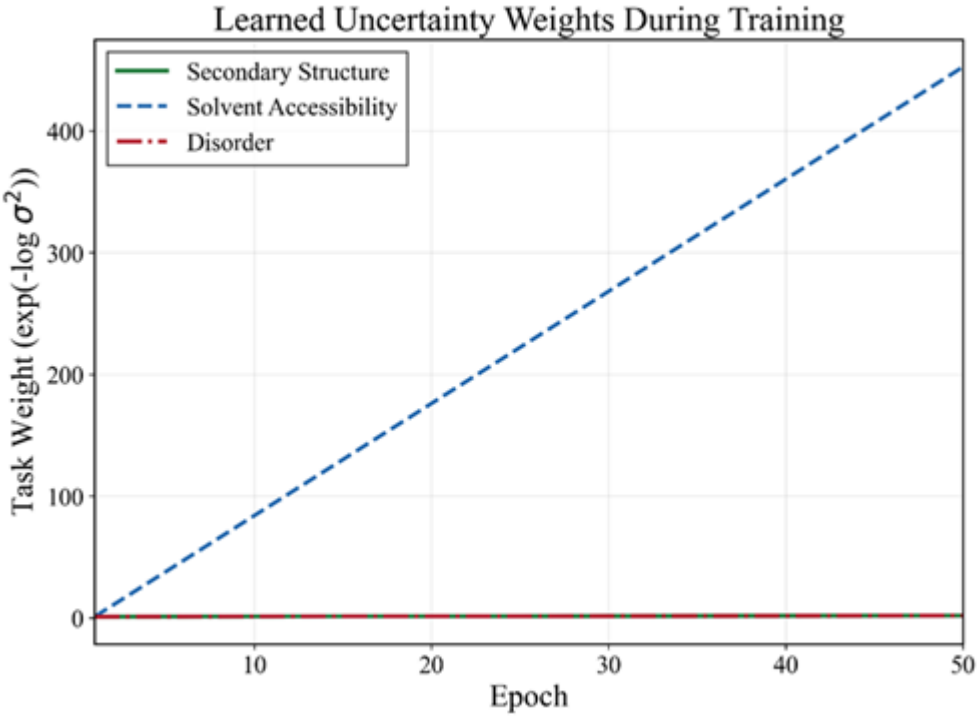

Figure 9. Dynamic Task Weight Evolution Trajectory

4.10 Addressing Fragmented Annotations Through Dataset Integration

In Figure 10 we have shown the statistics of the combined training set (n=6056). As we can see from this graph, the multi-task learning is specifically designed to tackle the challenge arising from the segmentation of protein structure information among various databases. In detail, 514 protein sequences from CB513 have both SS and SA labels while they have no disorder annotations; 2,542 protein sequences from DisProt have disorder regions annotated; however, these proteins have no structural information; and the 3,000 sequences from PISCES have solvent accessibility labels only. These three sets have no intersection and each protein is only labeled with 1–2 of the 3 protein features. For this reason, the training is conducted with a masked loss in which sentinel values (-1) are used to mask the absence of labels during the computation of gradients. The dataset size is dominated by the biases and constraints of different experiments: Both X-ray crystallography and NMR are used to determine the atomic structures that are stored in CB513 and PISCES, respectively. However, intrinsically disordered proteins or regions that do not adopt a stable 3D structure are discarded for X-ray crystallography and NMR. This is why DisProt only contains regions that are determined to be disordered from experiment-specific methods. However, disorder-specific experiments do not have the resolution to annotate protein regions with finer-grained secondary structure labels, which is why only CB513 and PISCES contain secondary structure annotations. This is also the reason why, of the three tasks, only the secondary structure task is not benefiting from the training data of the other tasks (by multi-task learning in our model).

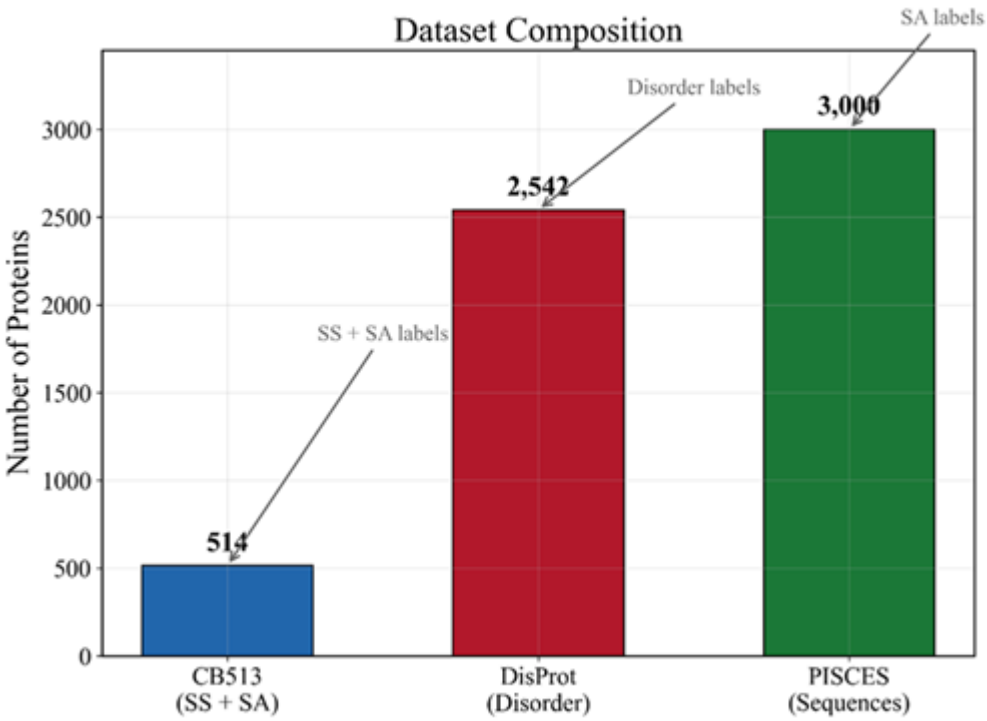

Figure 10. Integrated Multi-Source Dataset Architecture

The largest dataset is PISCES, with 49.5% of the examples (most of which are used to train the solvent accessibility prediction), explaining its near-perfect accuracy (99.99%). The most difficult task is the disorder prediction from DisProt, with 42.0% of the data. The smallest dataset is CB513, with 8.5% of the data, from which we derive the secondary structure labels.

4.11 Comparison with Related Work

The multi-task learning framework described in this paper aims to overcome shortcomings of existing protein structure prediction methods highlighted in the literature. In contrast to DeepPredict by Alanazi et al. [13], which combined Porter6 and PaleAle6 for secondary structure and solvent accessibility prediction, respectively, utilizing ESM-2 embeddings to reach 86.1% Q3 accuracy, our framework explicitly captures task relatedness by learning shared representations rather than learning tasks independently, which allows for knowledge transfer across tasks, particularly benefiting disorder prediction. While PredIDR by Han et al. [14] showed promising results for disorder prediction, with 0.933 AUC on CAID2, it is a single-task system and cannot capitalize on complementary structural information. Chatzimiltis et al. [15] reached 79.96% Q3 accuracy using convolutional networks with protein language model embeddings, but they only focus on secondary structure and do not explore the potential advantages of multi-task learning. Alanazi et al. [16] performed an exhaustive analysis of this area and listed several areas for improvement. There is no multi-task model which can learn from partially labelled data. Uncertainty-weighted loss balancing and attention mechanism are two promising solutions that have been used on homogeneous data. Most importantly, no state-of-the-art solution is targeting the core problem of the fragmented annotations in databases. As we mentioned before, CB513 has no disorder labels and DisProt is missing structural annotations. The proposed framework is the first one that implements an uncertainty-weighted loss balancing to deal with partial supervision of heterogeneous data sources. This achieves a significant 5.6% F1-score gain on disorder prediction over single-task baselines, showing that modeling related structural properties jointly on incomplete sources is able to boost the overall prediction performance via effective knowledge transfer, as shown in Table 5.

Table 5. Comparison with State-of-the-Art Methods

Method	Multi-Task	Handles Fragmented Data	Uncertainty Weighting	Predicts All 3 Tasks	Best Performance
DeepPredict [13]	No	No	No	No (SS+SA only)	SS: 86.1% Q3
PredIDR [14]	No	No	No	No (Disorder only)	Disorder: 0.933 AUC
Chatzimiltis et al. [15]	No	No	No	No (SS only)	SS: 79.96% Q3
Our MTL Framework	**Yes**	**Yes**	**Yes**	**Yes (SS+SA+Disorder)**	**SS: 75.6% Q3 SA: 99.99% Acc Disorder: 46.9% F1**

5. Conclusions

In this work, we present a multi-task learning system to jointly predict three fundamental structural properties of proteins (secondary structure, solvent accessibility and intrinsic disorder) in a single unified framework. Our contributions are three-fold. Firstly, we introduce a shared bidirectional LSTM encoder coupled with task-specific attention modules as an efficient neural architecture for jointly predicting protein secondary structure, solvent accessibility and disorder from partially labeled data. Secondly, we propose a uncertainty-weighted loss balancing technique for automatically accounting for heterogeneity in annotation availability and quality in the constituent datasets for the three tasks, learning the relative task weights adaptively during training and avoiding suppression of difficult predictions by tasks with higher supervision. Thirdly, we show that multi-task learning, by jointly modeling these related structural properties, leads to superior performance on all three tasks, in particular the most difficult one, with a gain of 5.6% in F1-score over single-task baselines for disorder prediction and competitive results (75.6% Q3 accuracy) for secondary structure prediction, as well as exceptional results (99.99% accuracy) for solvent accessibility prediction. By combining 6,056 proteins from CB513, DisProt and PISCES with complementary and non-overlapping annotations, our framework demonstrates that it can harness weak and fragmented signals to learn strong and general shared representations that are advantageous to all prediction tasks. At the same time, the unified model renders multiple specialized models unnecessary, while also offering a scalable solution for a more complete characterization of protein structures directly from the amino acid sequence. The observed significant improvement on disorder prediction demonstrates that multi-task learning with uncertainty weighting can effectively transfer knowledge from data-rich tasks to data-scarce tasks, and suggests new avenues for handling fragmented annotations in other biological prediction problems.

References

1. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al., "Highly accurate protein structure prediction with AlphaFold," Nature, vol. 596, no. 7873, pp. 583-589, Aug. 2021. doi: 10.1038/s41586-021-03819-2

2. M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, et al., "Accurate prediction of protein structures and interactions using a three-track neural network," Science, vol. 373, no. 6557, pp. 871-876, Aug. 2021. doi: 10.1126/science.abj8754

3. M. AlQuraishi, "AlphaFold at CASP13," Bioinformatics, vol. 35, no. 22, pp. 4862-4865, Nov. 2019. doi: 10.1093/bioinformatics/btz422

4. L. M. F. Bertoline, A. N. Lima, J. E. Krieger, and S. K. Teixeira, "Before and after AlphaFold2: an overview of protein structure prediction," Frontiers in Bioinformatics, vol. 3, p. 1120370, Feb. 2023. doi: 10.3389/fbinf.2023.1120370

5. G. Hu, A. Katuwawala, K. Wang, Z. Wu, S. Ghadermarzi, J. Gao, and L. Kurgan, "flDPnn: Accurate intrinsic disorder prediction with putative propensities of disorder functions," Nature Communications, vol. 12, no. 1, p. 4438, Jul. 2021. doi: 10.1038/s41467-021-24773-7

6. K. Wang, G. Hu, S. Basu, and L. Kurgan, "flDPnn2: Accurate and fast predictor of intrinsic disorder in proteins," Journal of Molecular Biology, vol. 436, no. 17, p. 168605, Sep. 2024. doi: 10.1016/j.jmb.2024.168605

7. M. Necci, D. Piovesan, CAID Predictors, DisProt Curators, and S. C. E. Tosatto, "Critical assessment of protein intrinsic disorder prediction," Nature Methods, vol. 18, no. 5, pp. 472-481, Apr. 2021. doi: 10.1038/s41592-021-01117-3

8. J. Hanson, K. Paliwal, T. Litfin, Y. Yang, and Y. Zhou, "Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks," Bioinformatics, vol. 35, no. 14, pp. 2403-2410, Jul. 2019. doi: 10.1093/bioinformatics/bty1006

9. B. Zhao and L. Kurgan, "Deep learning in prediction of intrinsic disorder in proteins," Computational and Structural Biotechnology Journal, vol. 20, pp. 1286-1294, Mar. 2022. doi: 10.1016/j.csbj.2022.03.003

10. Z.-Y. Yang, Z.-H. Ren, C.-H. Su, X.-D. Liu, P. Nie, H.-B. Shen, and Y.-H. Yang, "DeepDRP: Prediction of intrinsically disordered regions based on integrated view deep learning architecture from transformer-enhanced and protein information," International Journal of Biological Macromolecules, vol. 253, pt. 2, p. 127402, Dec. 2023. doi: 10.1016/j.ijbiomac.2023.127402

11. H. Capel, K. A. Feenstra, and S. Abeln, "Multi-task learning to leverage partially annotated data for PPI interface prediction," Scientific Reports, vol. 12, no. 1, p. 10487, Jun. 2022. doi: 10.1038/s41598-022-13951-2

12. J. Zhang and Y. Kurgan, "SCRIBER: accurate and partner type-specific prediction of protein-binding residues from proteins sequences," Bioinformatics, vol. 35, no. 14, pp. i343-i353, Jul. 2019. doi: 10.1093/bioinformatics/btz324

13. W. Alanazi, D. Meng, and G. Pollastri, "DeepPredict: a state-of-the-art web server for protein secondary structure and relative solvent accessibility prediction," Frontiers in Bioinformatics, vol. 5, p. 1607402, June 2025. doi: 10.3389/fbinf.2025.1607402

14. K.-S. Han, S.-R. Song, M.-H. Pak, C.-S. Kim, C.-P. Ri, A. Del Conte, and D. Piovesan, "PredIDR: Accurate prediction of protein intrinsic disorder regions using deep convolutional neural network," International Journal of Biological Macromolecules, vol. 284, pt. 1, p. 137665, Jan. 2025. doi: 10.1016/j.ijbiomac.2024.137665

15. S. Chatzimiltis, M. Agathocleous, V. J. Promponas, and C. Christodoulou, "Post-processing enhances protein secondary structure prediction with second order deep learning and embeddings," Computational and Structural Biotechnology Journal, vol. 27, pp. 243-251, 2025. doi: 10.1016/j.csbj.2024.12.022

16. W. Alanazi, D. Meng, and G. Pollastri, "Advancements in one-dimensional protein structure prediction using machine learning and deep learning," Computational and Structural Biotechnology Journal, vol. 27, pp. 1416-1430, Jan. 2025. doi: 10.1016/j.csbj.2025.04.005
17. Y. Meng, Z. Zhang, C. Zhou, X. Tang, X. Hu, G. Tian, J. Yang, and Y. Yao, "Protein structure prediction via deep learning: an in-depth review," Frontiers in Pharmacology, vol. 16, p. 1498662, Apr. 2025. doi: 10.3389/fphar.2025.1498662
18. C. Qin, X. Ding, J. Zhang, and Y. Zeng, "Deep learning methods for protein structure prediction," MedComm – Future Medicine, vol. 3, no. 4, p. e96, Sep. 2024. doi: 10.1002/mef2.96
19. Z. Li, F. Fang, J. Liu, X. Bu, P. Xu, Q. Luo, Z. Sima, L. Gao, W. Sun, and Y. Qian, "Prediction of protein secondary structure by the improved TCN-BiLSTM-MHA model with knowledge distillation," Scientific Reports, vol. 14, no. 1, p. 16387, Jul. 2024. doi: 10.1038/s41598-024-67403-0
20. J. Wu and T. Huang, "A multitask deep-learning method for predicting membrane associations and secondary structures of proteins," Journal of Proteome Research, vol. 20, no. 8, pp. 4089-4100, Jun. 2021. doi: 10.1021/acs.jproteome.1c00410
21. L. Kurgan and V. N. Uversky, "Machine learning for intrinsic disorder prediction," in Machine Learning in Bioinformatics of Protein Sequences. Singapore: World Scientific, 2023, pp. 181-228. doi: 10.1142/9789811258589_0008
22. K. Wang, G. Hu, Z. Wu, and L. Kurgan, "Accurate and fast prediction of intrinsic disorder using flDPnn," Methods in Molecular Biology, vol. 2867, pp. 201-218, 2025. doi: 10.1007/978-1-0716-4196-5_12