Stoic: Fast and accurate protein stoichiometry prediction

Daniil Litvinov
Biozentrum, University of Basel
SIB Swiss Institute of Bioinformatics
Basel, CHE
daniil.litvinov@unibas.ch

Lorenzo Pantolini
Biozentrum, University of Basel
SIB Swiss Institute of Bioinformatics
Basel, CHE
lorenzo.pantolini@unibas.ch

Peter Škrinjar
Biozentrum, University of Basel
SIB Swiss Institute of Bioinformatics
Basel, CHE
peter.skrinjar@unibas.ch

Gerardo Tauriello
Biozentrum, University of Basel
SIB Swiss Institute of Bioinformatics
Basel, CHE
gerardo.tauriello@unibas.ch

Caitlyn McCafferty
Biozentrum, University of Basel
Basel, CHE
caitlyn.mccafferty@unibas.ch

Benjamin D. Engel
Biozentrum, University of Basel
Basel, CHE
ben.engel@unibas.ch

Torsten Schwede
Biozentrum, University of Basel
SIB Swiss Institute of Bioinformatics
Basel, CHE
torsten.schwede@unibas.ch

Janani Durairaj
Biozentrum, University of Basel
SIB Swiss Institute of Bioinformatics
Basel, CHE
janani.durairaj@unibas.ch

Abstract

Protein structure prediction methods require prior knowledge of protein stoichiometry - the number of copies of each protein entity within a complex. Current approaches rely on computationally expensive brute-force methods that run structure prediction on multiple stoichiometry combinations, often with limited accuracy. We introduce *Stoic*, a method that uses protein language model embeddings to predict protein complex stoichiometry. Our approach learns to identify interface residues that participate in protein-protein interactions, rather than relying on global sequence features. By integrating these interface-aware embeddings into a graph neural network, *Stoic* achieves fast and accurate stoichiometry prediction for both homomeric and heteromeric targets.

1 Introduction

Protein structure prediction has undergone a revolutionary transformation with approaches such as AlphaFold2 (AF2) and others [1–4]. However, these breakthroughs have primarily focused on individual protein chains, while many biological processes depend on protein complexes composed of multiple subunits of the same or different protein entities. Often the information about which proteins assemble into a complex is known while stoichiometry - the number of copies of each unique protein chain is not. A critical limitation of current structure prediction methods is their requirement for this prior knowledge. Stoichiometry extends beyond individual structure prediction, since both major benchmarking efforts, CAMEO (Continuous Automated Model Evaluation) [5] and CASP16 (Critical Assessment of Structure Prediction) [6], have established stoichiometry prediction as the first task before complex structure modeling can proceed.

Current approaches to this challenge frequently rely on running AF2-like methods on multiple stoichiometry combinations, then using confidence scores to distinguish correct quaternary states [7, 8]. This brute-force approach is not only computationally expensive but also has limited accuracy, especially for large heteromeric targets [9]. Prediction methods to tackle related challenges for homomeric complexes are being developed, e.g. Seq2Symm [10] predicts protein symmetry, while QUEEN [11] predicts a copy number for homomers. These approaches demonstrate that protein language models (pLMs) can capture meaningful signals for stoichiometry-related tasks. However, they rely on average-pooled embeddings that capture global protein properties but often miss critical residue-specific nuances that are essential for understanding protein-protein interactions. In addition, prediction of stoichiometry of heteromeric complexes remains unsolved and less attempted.

Here we present *Stoic*, a novel method that combines graph neural networks with interface-specific protein language model embeddings to predict protein complex stoichiometry. *Stoic* learns to identify and weight interface residues, the specific amino acids that participate in protein-protein interactions, for embedding pooling rather than relying on global sequence features. By integrating these interface-aware embeddings into a graph neural network that introduces complex-related context, *Stoic* achieves fast and accurate stoichiometry prediction not only for homo- but also heteromeric targets.

2 Results

2.1 Graphs and interface-pooling improve stoichiometry prediction

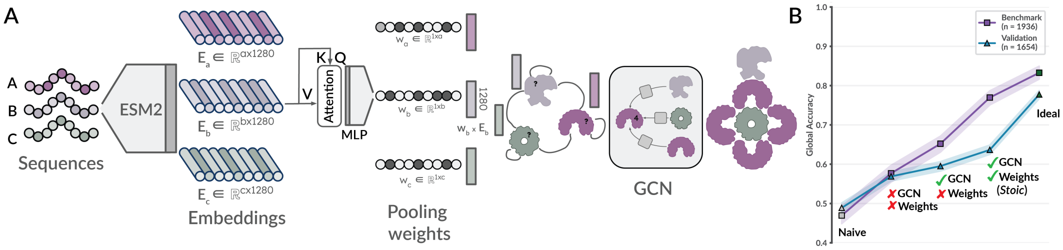


Figure 1: **Model and architectural choices. A)** Overview of the model architecture. **B)** Ablations removing different aspects of the network. *Naive* = predicting a copy number of 1 for all entities in a complex; *Ideal* = using the interface residues from ground truth complex structures to pool residue embeddings before the GCN step.

Architecture: Our model architecture integrates sequence embeddings from protein language models (pLMs) with graph neural networks to predict protein complex stoichiometry. As shown in Figure 1A, *Stoic* uses ESM2-650M [12] to obtain residue-level embeddings for each unique protein entity in a complex, and then uses a learned weighted pooling mechanism to aggregate residue embeddings into a fixed length protein embedding. The pooled embeddings act as node features in a fully connected graph, used as input to a graph convolutional neural network (GCN) [13] which outputs protein copy numbers as node labels. We formulate this as a multi-class classification problem, with 13 stoichiometry classes (1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 14, 24) representing 99.3% of the complexes in the Protein Data Bank (PDB) [14].

Dataset: The datasets used consist of biological assemblies from the PDB [14] after applying various filtering criteria described in Supplementary Table S1. The training and validation sets consist of complexes with a release date before 30 September 2021. We used MMseqs2 [15] clustering with a minimum of 30% sequence identity and 80% coverage to assign a cluster label to each protein entity within each complex, and then define the complex cluster label as the combination of the entity cluster labels. Training and validation sets thus have distinct complex cluster labels, ensuring that proteins with high sequence similarity are assigned to the same split, preventing data leakage. The benchmark dataset consists of all complexes released after 30 September 2021 where at least one entity has <30% sequence identity to any complex in the training set. To eliminate redundant entries, we sorted the data by resolution and removed duplicates based on the combination of sequences and corresponding stoichiometry, retaining the structure with the highest resolution for each unique combination. For each complex, we extracted the copy number of each unique entity from its first biological unit. The combination of copy numbers for a complex (stoichiometry) served as our target variable. For training the weighted pooling head, we also extract interface residues from each entity as any residue with any atom within 8Å from another protein in the complex.

Losses: The model is trained using two complementary losses. The first (1) is the main complex loss for the copy number classification task, which steers the model to predict global stoichiometry correctly. This loss uses the sum of logarithms of individual entity cross-entropy losses between l_i (logits of individual entity) and c_i (copy number class) across all n entities in a complex. This is mathematically equivalent to multiplication but provides better gradient properties during training. This formulation penalizes the model more heavily for incorrectly predicting any single component of a complex compared to standard cross-entropy loss, where such errors would be obscured, especially for complexes with many entities. To counteract the effects of extreme class imbalance (e.g. class 7 and 9 make up 0.3% of the training set) we introduced class weights calculated using effective number of samples defined as $(1 - \beta^{n_i})/(1 - \beta)$, where n_i is the number of samples of class i and β = 0.9999 is a hyperparameter [16]. The second (2) is an auxiliary focal loss for predicting interface residues (r), which helps the model learn appropriate weights (w) for the pooling mechanism by identifying which residues are most relevant for stoichiometry prediction.

$$\mathcal{L}_{\text{complex}} = \sum_{i=1}^{n} \ln\left(\mathcal{L}_{\text{CE}}(l_i, c_i) + 1\right) \tag{1}$$

$$\begin{gathered}\mathcal{L}_{\text{interface}} = \sum_{i=1}^{n} \alpha_i (1 - w_i')^2 \mathcal{L}_{\text{BCE}}(w_i, r_i) \\ \text{where } \alpha_i, w_i' = \begin{cases} 0.75, w_i & \text{if } r_i = 1 \\ 0.25, 1 - w_i & \text{if } r_i = 0 \end{cases}\end{gathered} \tag{2}$$

The overall stoichiometry for a complex is predicted as follows: for each entity, the model outputs logits for the 13 copy number classes, which are converted to probabilities using a *softmax* operation. Top-N stoichiometry candidates are generated by computing the Cartesian product of the highest probability values across all proteins in the complex and ranking these combinations by their joint probability. *Stoic* was trained for 64 epochs (based on early stopping) with starting learning rate 3e-4, AdamW optimiser with weight decay 0.05, and the ReduceLROnPlateau learning rate scheduler.

We show the power of this architecture through baselines and ablations in Figure 1B, with the global stoichiometry prediction accuracy on cluster representatives from the validation set (in blue) and the benchmark set (in purple). Our Naive baseline simply predicts a copy number of one for all entities in the complex, the current approach used in CAMEO [5] for their AlphaFold3 [4] baseline. This is followed by an MLP on average-pooled embeddings i.e. no graph and no weighted pooling. This approach is used by models such as QUEEN [11] and Seq2Symm [10] for related tasks, and thus acts as a proxy for the performance of these methods on this task. Adding the context of the other proteins in the complex through graph convolutions improves prediction performance, and further adding learned weighted pooling using an interface-specific auxiliary loss reaches 30 percentage points higher accuracy compared to the baseline. We also show the idealistic scenario where interface residues are obtained from the ground truth structure and used for pooling, demonstrating that interface residues contain very relevant and strong signals for predicting stoichiometry, and supporting our approach to employ auxiliary interface losses to improve pooling and thus prediction performance.

2.2 *Stoic* provides accurate and generalisable stoichiometry predictions

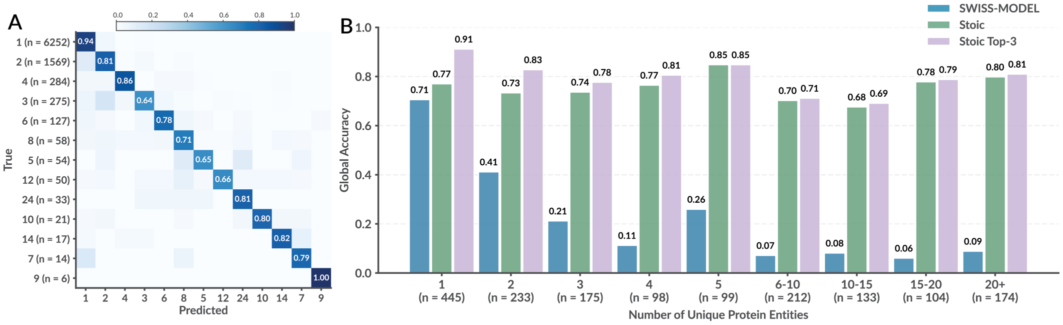


Figure 2: **Copy number and stoichiometry prediction performance. A)** Confusion matrix of *Stoic* copy number prediction across all proteins in the Benchmark set with <30% sequence identity to the training set. **B)** Global stoichiometry prediction accuracy across the Benchmark set for SWISS-MODEL (blue), *Stoic* (green), and *Stoic* Top-3 (purple), divided across complexes with differing number of unique entities.

We first look at node-level classification performance for copy number prediction. The confusion matrix for this task only on proteins with <30% sequence identity to our training set (Figure 2 A) shows good performance across the board, including rare stoichiometries (e.g. 7 and 9), with only some discrepancies for trimers versus dimers.

This generalisability also holds true for global stoichiometry prediction on the benchmark dataset, where a prediction is labeled correct only when all the individual protein entities within a complex have their copy numbers predicted correctly (Figure 2B). For this global task, we also compare to a template-based approach SWISS-MODEL [17] which uses HHSearch [18] against the PDB (with a release date before 30 September 2021) followed by QSQE-based template selection [19]. We consider SWISS-MODEL prediction as a success if any of the selected templates has the same stoichiometry as the target. *Stoic* outperforms SWISS-MODEL in assigning the correct stoichiometry, especially in complexes with many individual entities where finding templates containing homologs for all entities becomes difficult.

The purple bars, which show the accuracy if any of the top 3 predicted stoichiometries are correct, shows that further room for improvement in stoichiometry ranking remain and also potentially allow for cases where a protein differs in its stoichiometry based on its environment, for example, the metamorphic protein Selecase [20].

2.3 Embeddings from selected interface residues inform complex stoichiometry

Stoic learns to predict residues participating in protein interfaces as a secondary task, using these intermediate predictions to generate more meaningful embeddings for the main stoichiometry prediction task. We showcase some examples of these residue predictions in Figure 3A, where residues with an assigned weight of >0.4 for one protein in each complex are highlighted. These results are also supported by the precision-recall curve for predicting interface residues shown in Figure 3B, demonstrating good performance on this task. It is interesting to note that not all interface residues are relevant for stoichiometry prediction and, conversely, not all residues with high weights are in the interface. Further exploration of the residues selected by *Stoic* could be useful for interpretability studies exploring complex formation.

3 Discussion

Our results demonstrate that *Stoic* successfully addresses the critical bottleneck of stoichiometry prediction in determining protein complex structure. The design of our model makes it easy to integrate with existing structure prediction pipelines - *Stoic* can serve as initial component providing

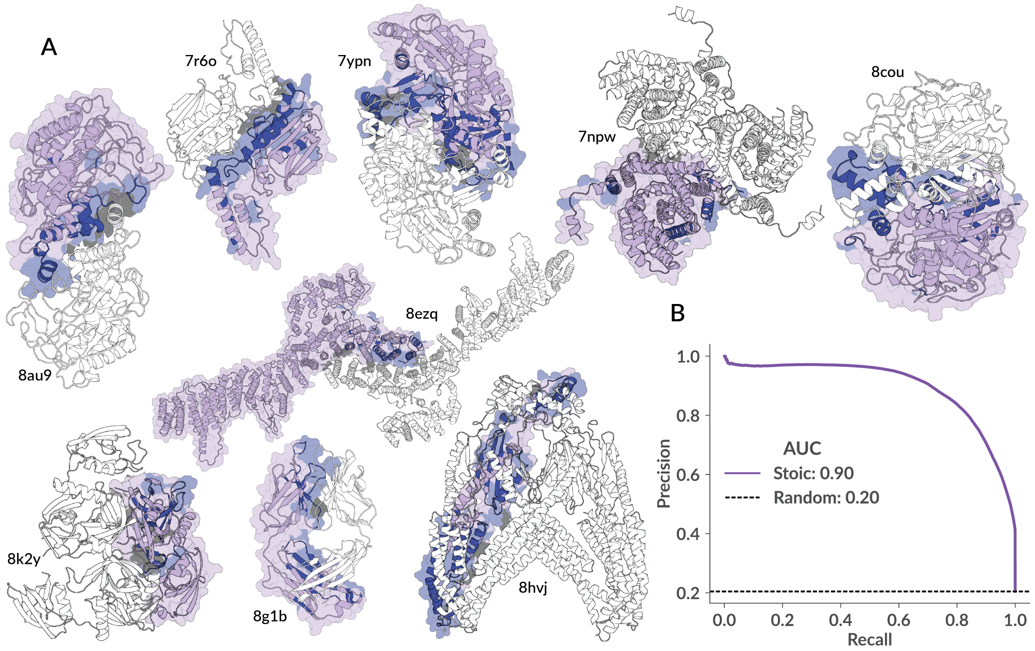


Figure 3: **Interface residues relevant for stoichiometry prediction. A)** Examples of *Stoic* residue weights for one chain each of complexes where all entities have <30% sequence identity to the training set. For each complex, residues with predicted weights >0.4 for one chain are highlighted in blue. **B)** The precision-recall curve for interface residue prediction on the Benchmark dataset, shown for *Stoic* (purple) as well as for a baseline (dashed black) the fraction of interface residues. The optimal operating point of the *Stoic* PR curve determined the 0.4 threshold for A.

a set of the most likely stoichiometries for multimeric structure prediction methods such as AlphaFold-Multimer [21], AlphaFold3 [4] etc.

There are several promising directions for future development. First, incorporating negative examples during training could further improve model performance. Currently, our training data consist only of positive examples of known protein complexes. Adding negative examples, such as proteins that are known not to interact or form stable complexes, could help the model better distinguish between plausible and implausible stoichiometries. This would be particularly valuable for identifying cases in which proteins might form different complexes under different conditions or where some parts of a complex are unknown. The top-3 global accuracy results suggest that there remains room for improvement in stoichiometry ranking. This could be addressed through more sophisticated ranking algorithms, i.e. by implementing a confidence prediction head.

The interface-aware pooling mechanism developed in *Stoic* has broader implications beyond stoichiometry prediction. Residue-level pooling that focuses on functionally relevant regions could be valuable for many other protein-related tasks, including protein-protein interaction prediction, functional site identification, and drug binding site prediction. The ability to learn which residues are most important for a given task through auxiliary losses represents a generalizable approach that could be applied across diverse problems.

References

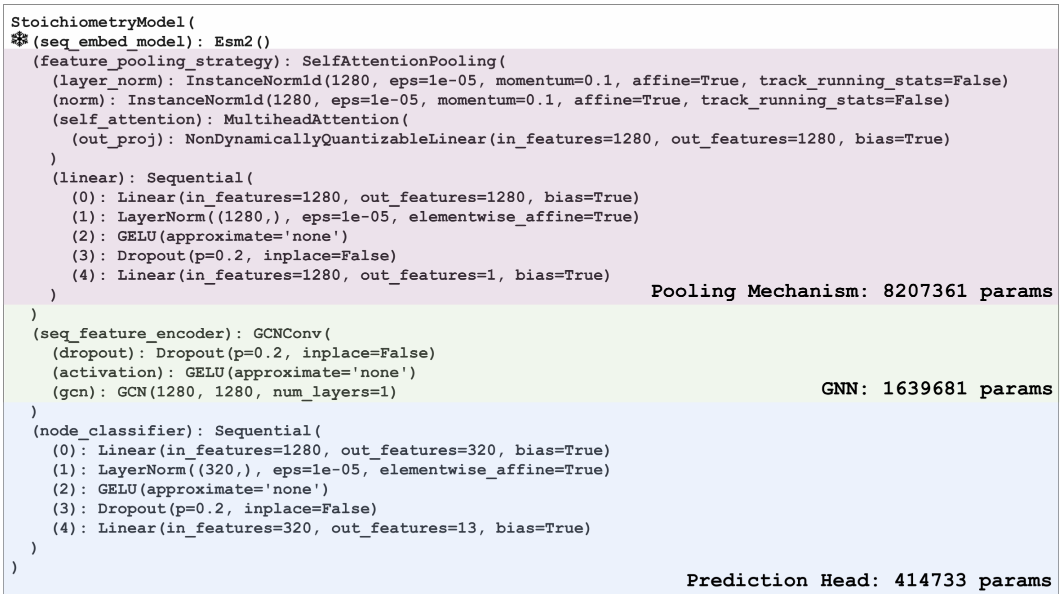
1. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *nature* **596,** 583–589 (2021).
2. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373,** 871–876 (2021).
3. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379,** 1123–1130 (2023).

4. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630,** 493–500 (2024).
5. Robin, X. *et al.* Benchmarking of macromolecular complexes with the Continuous Automated Model Evaluation (CAMEO). *Authorea Preprints* (2025).
6. Zhang, J. *et al.* Assessment of Protein Complex Predictions in CASP16: Are we making progress? *bioRxiv,* 2025–05 (2025).
7. Shor, B. & Schneidman-Duhovny, D. CombFold: predicting structures of large protein assemblies using a combinatorial assembly algorithm and AlphaFold2. *Nature methods* **21,** 477–487 (2024).
8. Liu, J., Neupane, P. & Cheng, J. Accurate prediction of protein complex stoichiometry by integrating alphafold3 and template information. *bioRxiv* (2025).
9. Elofsson, A. AlphaFold3 at CASP16. *Proteins: Structure, Function, and Bioinformatics* (2025).
10. Kshirsagar, M. *et al.* Rapid and accurate prediction of protein homo-oligomer symmetry using Seq2Symm. *Nature Communications* **16,** 2017 (2025).
11. Avraham, O., Tsaban, T., Ben-Aharon, Z., Tsaban, L. & Schueler-Furman, O. Protein language models can capture protein quaternary state. *BMC bioinformatics* **24,** 433 (2023).
12. Rives, A. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* **118,** e2016239118 (2021).
13. Kipf, T. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* (2016).
14. Burley, S. K. *et al.* Protein Data Bank (PDB): the single global macromolecular structure archive. *Protein crystallography: methods and protocols,* 627–641 (2017).
15. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology* **35,** 1026–1028 (2017).
16. Cui, Y., Jia, M., Lin, T.-Y., Song, Y. & Belongie, S. *Class-balanced loss based on effective number of samples* in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019), 9268–9277.
17. Waterhouse, A. *et al.* SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic acids research* **46,** W296–W303 (2018).
18. Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *BMC bioinformatics* **20,** 473 (2019).
19. Bertoni, M., Kiefer, F., Biasini, M., Bordoli, L. & Schwede, T. Modeling protein quaternary structure of homo-and hetero-oligomers beyond binary interactions by homology. *Scientific reports* **7,** 10480 (2017).
20. López-Pelegrín, M. *et al.* Multiple stable conformations account for reversible concentration-dependent oligomerization and autoinhibition of a metamorphic metallopeptidase. *Angewandte Chemie* **126,** 10800–10806 (2014).
21. Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. *biorxiv,* 2021–10 (2021).

A Technical Appendices and Supplementary Material

Supplementary Table S1: **Dataset Statistics.** Number of protein complexes at different stages of filtering.

Filtering Step	**№ Complexes**
Original PDB	225,355
Remove structures without resolution	212,368
Remove capsids	195,655
Remove proteins with helical symmetry	195,607
Remove non-protein structures	194,608
Sequence length filtering	194,404
Deduplication	112,924



```
StoichiometryModel(
❄(seq_embed_model): Esm2()
  (feature_pooling_strategy): SelfAttentionPooling(
    (layer_norm): InstanceNorm1d(1280, eps=1e-05, momentum=0.1, affine=True, track_running_stats=False)
    (norm): InstanceNorm1d(1280, eps=1e-05, momentum=0.1, affine=True, track_running_stats=False)
    (self_attention): MultiheadAttention(
      (out_proj): NonDynamicallyQuantizableLinear(in_features=1280, out_features=1280, bias=True)
    )
    (linear): Sequential(
      (0): Linear(in_features=1280, out_features=1280, bias=True)
      (1): LayerNorm((1280,), eps=1e-05, elementwise_affine=True)
      (2): GELU(approximate='none')
      (3): Dropout(p=0.2, inplace=False)
      (4): Linear(in_features=1280, out_features=1, bias=True)
    )                                                    Pooling Mechanism: 8207361 params
  )
  (seq_feature_encoder): GCNConv(
    (dropout): Dropout(p=0.2, inplace=False)
    (activation): GELU(approximate='none')
    (gcn): GCN(1280, 1280, num_layers=1)                                GNN: 1639681 params
  )
  (node_classifier): Sequential(
    (0): Linear(in_features=1280, out_features=320, bias=True)
    (1): LayerNorm((320,), eps=1e-05, elementwise_affine=True)
    (2): GELU(approximate='none')
    (3): Dropout(p=0.2, inplace=False)
    (4): Linear(in_features=320, out_features=13, bias=True)
  )
)
                                                            Prediction Head: 414733 params
```

Supplementary Figure S1: **Detailed architecture of *Stoic*.**