THE UNIVERSITY of EDINBURGH
School of Biological Sciences

COVER SHEET

Dissertation Title
Predicting Novel Enzymes in Trypanosomes using Unsupervised Learning and Molecular Docking

Student Exam Number:	B200694

Submitted in partial fulfilment of the requirements for the Degree of Master of Science in
Bioinformatics
of the University of Edinburgh 2024-2025

Name(s) of Supervisor(s):
Dr. Chris Wood, Dr. Bungo Akiyoshi

Declaration	
I **have/have not** used Generative AI-based tools in the preparation of this Dissertation	I have
If you have: please provide the name and version of **ALL** GenAI tools used (e.g. Grammarly, Co-pilot, ELM, ChatGPT-3.5, o3, Llama 3.3, Gemini, etc.)	GPT-4o, GPT-5, Claude-4o-sonnet
URL links to **each** of the GenAI tools	https://chatgpt.com/, https://claude.ai/
Descriptions of contexts in which **each** of the tools were used	Claude for code generation and debugging. GPT for general scientific discussion, and for recommending changes to the text to improve clarity, coherence, brevity and register.

Word Count*	9030

Abstract:

Protein sequence discovery has accelerated dramatically in recent years. However, experimental functional characterisation—especially in non-model organisms—has not kept pace. Leveraging recent innovations in deep learning, this study presents a computational framework that narrows large, poorly annotated proteomes to a small set of candidate proteins with a shared, hypothesised function that can be prioritised for experimental follow-up. We design this pipeline to identify putative histone methyltransferases in *Trypanosoma brucei*, a kinetoplastid parasite responsible for vector-borne human and animal diseases. In particular, we aim to find an enzyme responsible for N-terminal methylation of histone H3, which has been implicated in chromosome segregation fidelity and may represent a trypanosome-specific regulatory mechanism. Using Evolutionary Scale Modelling (ESM-1b/ESM-2) to generate proteome-wide embeddings, t-distributed stochastic neighbour embedding (t-SNE) was applied to visualise structural, evolutionary, and functional similarity against curated methyltransferase reference datasets. Candidates were selected based on localisation to reference clusters and structurally modelled with Boltz-2 in complex with histone H3 and S-adenosyl methionine (SAM), the predicted methyl donor. Ligand binding was evaluated with GNINA, and catalytic targets predicted from binding site geometry. Of 61 initial candidates, 51 met affinity thresholds for predicted SAM-binding activity, with 13 predicted as competent for methylation of histone H3 residues. The final ranked list was enriched for proteins with putative annotations supporting the biological plausibility of predictions. Importantly, we nominate one potential N-terminal methyltransferase for high-priority experimental follow-up. This work demonstrates that combining unsupervised representation learning with co-folding and docking offers a scalable and flexible framework for guiding experimental functional validation in poorly annotated proteomes across the tree of life.

Table of Contents:

0. Table of Abbreviations 4
1. Introduction: 5
1.1 Emerging methods for protein function prediction: 6
1.2 Biological rationale: 8
1.3 A novel approach to protein identification: 10
2. Methodology: 11
2.1 Dataset generation: 11
2.2 Unsupervised learning: 11
2.2.1 Embedding generation with Evolutionary Scale Modelling (ESM): 11
2.2.2 Dimensionality Reduction with t-SNE: 12
2.2.3 Functional annotation with eggNOG-mapper: 13
2.2.4 Evaluation of Unsupervised learning methods: 14
2.2.5 Candidate selection: 15
2.3 Structure-based molecular docking: 15
2.3.1 Co-folding with Boltz-2: 15
2.3.2 Binding affinity prediction with GNINA: 16
2.4 Candidate filtering and ranking: 17
3. Results: 19
3.1 Evaluation of t-SNE performance: 19
3.2 Evaluation of ESM and t-SNE with respect to biological significance: 20
3.3 Candidate selection: 24
3.4 Comparison of Boltz-2 and GNINA affinity predictions: 25
3.5 Binding conditions: 27
3.6 Predicting SAM-binders: 28
3.7 Predicting histone H3 methyltransferases: 28

4. Discussion: 32
4.1 On Unsupervised learning: 32
4.2 On candidate selection: 35
4.3 On structure-based molecular docking: 36
4.4 Nominated methyltransferases and future work 37
5. Conclusion: 41
6. References: 42
7. Supplementary Materials 52
8. Acknowledgements: 55

0. Table of Abbreviations

Abbreviation:	Meaning:
CNN	Convolutional neural network
COG	Clusters of Orthologous Genes
EC	Enzyme Commission
ESAG	Expression site associated gene
ESM	Evolutionary Scale Modelling
KEGG	Kyoto Encyclopedia of Genes and Genomes
PLM	Protein language model
RHS	Retrotransposon hot spot
SAM	S-adenosyl L-methioinine
t-SNE	t-Distributed stochastic neighbour embedding
VSG	Variable surface glycoprotein

1. Introduction:

Central to nearly all biological processes, proteins are the driving forces of life. Characterising proteins in terms of their sequence, structure, and function, as well as the biological context in which they operate, are essential prerequisites in the development of novel treatments against disease. However, while the quantity of protein sequence data has grown exponentially in recent decades, the availability of experimentally validated information attached to those sequences is lagging behind. As of August 2025, the UniProt Knowledgebase contains over 253 million protein sequences, but under 0.23% of them have manually reviewed annotations (The UniProt Consortium, 2025). The challenge of distinguishing protein function is particularly acute in non-model organisms, where experimental characterisation is rarer still.

Trypanosoma brucei—a parasitic kinetoplastid responsible for African trypanosomiasis (sleeping sickness) in humans and nagana in livestock (Steverding, 2016)—is one such organism. Its highly divergent proteome, non-canonical biology, and disease burden make it a useful test case for developing scalable and labour-efficient computational pipelines that can be translated across species. Here, the specific problem is to identify an uncharacterised enzyme responsible for modifying *T. brucei* histones. Given its hypothesised role in cell division and possible trypanosome-specific lineage, the enzyme has been proposed as a candidate drug target. While most computational methods are optimised to infer biological function from a known amino acid sequence, here the challenge is reversed; the function is known, but the sequence is not. Nevertheless, by leveraging comparative datasets of functionally similar enzymes in other organisms, many of the same tools remain applicable. To address this task, it is first necessary to consider recent advances in computational protein function prediction that can be adapted to infer candidate sequences from a defined function. Below, we review key developments in this field, before returning to the biological rationale for targeting the uncharacterised *T. brucei* enzyme and outlining our chosen approach.

1.1 Emerging methods for protein function prediction:

The guiding principles of protein function prediction are united under the sequence-structure-function paradigm, which contends that a protein's amino acid sequence dictates its structure, and its structure determines its function (Anfinsen, 1973). With the creation of protein sequence databases, therefore, early functional annotations relied on sequence-based approaches such as BLAST, which identifies homologous sequences through alignment and infers functional information from well-characterised matches (Altschul et al., 1990). However, proteins with similar sequences can have highly divergent functions, especially in the case of paralogous protein families arising from gene duplication events (Hittinger and Carroll, 2007). Conversely, proteins with similar functions can evolve from different ancestral origins via convergent evolution and therefore lack detectable sequence similarity (Chen, DeVries and Cheng, 1997). Sequence-based approaches are therefore limited for all but close homologs with highly conserved sequences.

Moving one level up the paradigm, the notion that structure is a better predictor of function than sequence has become consensus in the scientific community (Avery et al., 2022). In this context, deep learning has made a significant impact. For example, DeepFRI and HEAL each use different neural network architectures to infer gene ontology (GO) terms from protein structures (Gligorijević et al., 2021; Zhonghui et al., 2023), and while experimental determination of protein structures remains costly and time-consuming, these tools can utilise AlphaFold-predicted structures to improve accuracy and accessibility (Jumper et al., 2021; Baek et al., 2021; Ma et al., 2022; Zhonghui et al., 2023). These methods have transformed the landscape of protein function prediction. However, AlphaFold is typically optimised for cases where high-quality multiple sequence alignments (MSAs) and structural templates are available (Gomez and Kovalevskiy, 2024), and in many non-model organisms such information can be sparse.

Protein language models (PLMs) offer a complementary route to function prediction. Much like language models in natural language processing learn grammar and meaning from large collections of text, protein language models are trained on vast

numbers of protein sequences to learn the ‘grammar’ and ‘vocabulary’ of proteins—the statistical patterns and relationships that underpin their structure and function. They can operate directly on single sequences without requiring MSAs or template structures, run at higher speeds, and generalise well across poorly characterised taxa (AlphaFold and beyond, 2023). This flexibility has conferred applications across multiple disciplines—including protein design, drug discovery, and genomics—as well as in protein function prediction (He et al., 2024; Zheng et al., 2024; Hu et al., 2024; Unsal et al., 2022). Among these, the Evolutionary Scale Modelling (ESM) suite of PLMs has attracted significant attention for its accuracy, speed, and range of applications (Rives et al., 2021; Lin et al., 2023; Chen et al., 2025). Built on the Transformer architecture (Vaswani et al., 2017), ESM models are adept at identifying complex patterns and long-range dependencies spanning entire protein sequences. Given that protein sequences are shaped by evolution to preserve structure and function, ESM can implicitly learn biochemical, physical, structural and functional properties from sequence alone. It outputs these properties as embeddings—numerical, vectorised representations of the input sequence—which can then be applied to various supervised and unsupervised machine learning tasks. For example, an ESM-1b-based variant effect prediction pipeline outperformed state-of-the-art methods in distinguishing pathogenic from benign variants in benchmark human genome datasets (Brandes et al., 2023). Overall, ESM’s capacity to unify diverse signals and information types makes it a promising complement to structure-based functional prediction, particularly for proteins lacking close structural or sequence homologs. Furthermore, the ability to cluster functionally similar enzymes based on embedding profiles (Yeung et al., 2023) makes it suitable to our task of identifying proteins of known function but unknown sequence.

Resolving the molecular partners of proteins, from nucleic acids to small-molecule cofactors, ions, or other proteins, is another pillar of functional prediction—but one not considered by structure-based methods or protein-language models alone. To this end, co-folding software like AlphaFold3, Chai-1, and Boltz-1, have substantially elevated our ability to predict the structural conformations of interacting biomolecules (Rennie and Oliver, 2025), while molecular docking software such as AutoDock and GNINA can characterise ligand binding modes and affinities (Forli et al., 2016; McNutt et al., 2025). Indeed, AutoDock has been successfully applied to ‘one substrate many

enzymes' virtual screening tasks to uncover novel genes in mammals (Malatesta et al., 2024). Boltz-2, released in June 2025, is the latest of this new generation of tools and can jointly predict binding affinities as well as 3D structures from training on experimental structures, molecular dynamics ensembles, and standardized biochemical assay measurements (Passaro et al., 2025). As a recent tool, Boltz-2 is relatively untested across diverse protein families and biological contexts. Its novel capabilities, however, make it an intriguing candidate to evaluate in the identification of functionally relevant protein interactions.

The integration of multiple information sources—such as sequence data, structural models, protein interactions, and domain annotations—has emerged as a powerful strategy and a major trend in the field of functional prediction (Lin et al., 2024). Together, PLMs and co-folding software can expand functional prediction beyond the limits of any single paradigm and have made the task of predicting novel proteins *in silico* more feasible than ever, especially in non-model organisms such as *Trypanosoma brucei.* We contend that leveraging ESM to cluster functionally similar proteins, followed by co-folding to model biomolecular interactions, could identify proteins where a known sequence is absent.

1.2 Biological rationale:

In *T. brucei*, as in most eukaryotes, chromosome segregation is orchestrated by the kinetochore, a proteinaceous structure that assembles on centromeres and connects them to spindle microtubules during mitosis (Akiyoshi and Gull, 2014). Typically, kinetochore assembly sites are epigenetically marked by the presence of CENP-A, a variant of histone H3 (Westhorpe and Straight, 2013). However, Trypanosomes both lack this variant and possess a highly divergent set of kinetochore proteins—none of which are canonical in other eukaryotic systems (Ishii and Akiyoshi, 2020). As a result, the mechanisms of kinetochore localization remain largely undefined in Trypanosomes, but the proteins involved represent attractive drug targets (Ishii and Akiyoshi, 2020).

Post-translational modifications (PTMs) are key epigenetic regulators of kinetochore assembly. In particular, methylation of the CENP-A N-terminal by the NTMT1 enzyme is essential for recruiting specific kinetochore components and for maintaining chromosome segregation fidelity (Sathyan, Fachinetti and Foltz, 2017). The biological significance of N-terminal methylation appears to extend to *T. brucei*; preliminary studies of the *T. brucei* subspecies *T. b. brucei*, led by Dr. Bungo Akiyoshi at the University of Edinburgh, have identified a kinetochore protein that specifically binds histone H3 when the serine residue at its N-terminus (H3-Ser-1) is unmethylated. However, the X-Pro-Lys consensus sequence of both NTMT1 and NTMT2 is absent from the *T. b. brucei* histone H3 tail, suggesting that is not a substrate for either of the only known eukaryotic N-terminal methyltransferases (Jakobsson et al., 2018; Deák et al., 2023). It is therefore hypothesized that H3-Ser-1 methylation is mediated by an uncharacterised enzyme. Given the apparent importance of the H3-Ser-1 methylation status for kinetochore protein binding, this unknown methyltransferase represents a compelling target for characterisation—both to elucidate the unique epigenetic regulation in trypanosomes and as a potential drug target.

While the primary aim of this study is to identify the specific methyltransferase responsible for modifying the N-terminus of histone H3 in *T. b. brucei*, any broader utility should not be overlooked. Methyltransferases constitute a diverse family of enzymes that catalyse the transfer of methyl groups—typically from the donor molecule S-adenosyl L-methionine (SAM)—to a wide range of biological substrates including DNA, RNA, and proteins (Wong and Eirin-Lopez, 2021). Beyond kinetochore assembly, they are central to the epigenetic and epitranscriptomic regulation of many biological processes (Wong and Eirin-Lopez, 2021). In Trypanosomes, several methyl marks have been identified at sites of transcription initiation (Kraus et al., 2020). Importantly, the number of modifications exceeds what would be expected from the small number of known histone methyltransferases—implying that additional, yet unidentified, enzymes remain to be discovered. N-terminal methylation is comparatively understudied but is known to be varied and context-dependent (Diaz, Meng and Huang, 2021). At the molecular level, it can influence protein localisation, stability, and degradation; more broadly, it affects protein-protein and protein-DNA interactions, thereby playing roles in mitosis, chromatin organisation, tRNA trafficking, DNA repair, and genome maintenance (Diaz, Meng and Huang, 2021). In this context,

a recent study by Yeung et al. (2023) demonstrated that embeddings from the ESM-1b protein language model can be used to effectively cluster functionally similar SAM-binding enzymes. Therefore, leveraging ESM to identify multiple novel methyltransferases in *T. b. brucei* may shed light on fundamental regulatory mechanisms and reveal new therapeutic targets.

1.3 A novel approach to protein identification:

Typically, functional prediction begins with an amino acid sequence from which biological function is inferred—such as in the Critical Assessment of Function Annotation (CAFA) community project (Zhou et al., 2019). However, given the lack of a known sequence for the target N-methyltransferase in *T. brucei*, we instead designed a novel multi-step computational workflow that combines ESM to cluster and identify functionally similar enzymes, Boltz-2 structural prediction for modelling protein–SAM and protein-histone interactions, and molecular docking to assess binding specificity. Together, this approach enabled the nomination of several candidate methyltransferases and their methylation targets, including one which may mediate H3 N-terminal methylation in *T. b. brucei*.

2. Methodology:

2.1 Dataset generation:

To generate a dataset representing the complete proteome of *Trypanosoma brucei brucei* strain TREU927, sequences were obtained from both UniProt and NCBI Protein databases (Bateman et al., 2024; Sayers et al., 2024). Query results from each database were combined and identical sequences were removed in Biopython v1.85, yielding 8937 total sequences. Reference datasets comprised of reviewed UniProt protein sequences were organised into four tiers by Gene Ontology: Tier 1 proteins show N-terminal protein N-methyltransferase activity (go:0071885), Tier 2 proteins show histone H3 methyltransferase activity (go:0140938), Tier 3 proteins show histone methyltransferase activity (go:0042054), and Tier 4 proteins show protein methyltransferase activity (go:0008276). All reference datasets were further filtered for SAM-binding activity (CHEBI:15414). To minimize redundancy that may skew downstream dimensionality reduction, reference datasets were processed with CD-HIT v4.8.1 using an identity threshold of 90%—calculated from the number of shared short substrings (e.g., dipeptides or tripeptides) between sequences—and then clustered based on their similarity to the representative (longest) sequence in each cluster (Li and Godzik, 2006). Only the reference sequence from each cluster was retained, yielding 13, 249, 363, and 480 sequences for each of tiers 1-4, respectively.

2.2 Unsupervised learning:

2.2.1 Embedding generation with Evolutionary Scale Modelling (ESM):

Since classical machine learning algorithms operate on numerical input data, protein sequences must first be converted into a format suitable for use as feature representations, such as vectorised embeddings. To achieve this, we used pretrained ESM-1b and ESM-2 transformer PLMs (Rives et al., 2021; Lin et al., 2023). Both

models are trained using a masked language modelling (MLM) approach that minimises the following loss function:

$$\mathcal{L}_{\text{MLM}} = -\sum_{i \in M} \log p\left(x_i | x_{\setminus M}\right)$$

In this objective, a subset of positions M in the protein sequence are randomly masked, and the model is trained to predict the original amino acid x_i at each marked position i, given the rest of the sequence $x_{\setminus M}$ as context. This loss function encourages the model to learn the statistical dependencies and contextual relationships between amino acids in the sequence, enabling it to build meaningful internal embeddings capturing protein structure and function.

FASTA files were processed in fair-esm v2.0.1 using the esm1b_t33_650M_UR50S and esm2_t33_650M_UR50D models, outputting a 1028 x 1 vector in PyTorch file format for each sequence, with each dimension corresponding to the mean embedded value across all residues. At this stage, tiers 2 and 4 of the reference datasets were discarded to reduce computational expense. Tier 1 was retained as N-terminal N-methyltransferases were considered closest in function to the target protein, while tier 3 was retained to capture additional proteins capable of binding the histone tail.

2.2.2 Dimensionality Reduction with t-SNE:

Due to the Transformer architecture underlying ESM models, the meaning of individual dimensions within the embedding space is abstract and not directly interpretable. Nevertheless, we contend that proximity between embeddings at the global level should reflect functional similarity. To visualise these relationships, we performed t-distributed stochastic neighbour embedding (t-SNE) on protein embeddings and plotted the position of each protein in 2D. t-SNE is a non-linear dimensionality reduction algorithm which models high-dimensional pairwise similarities as probabilities and embeds data in a lower-dimensional space by minimizing the Kullback–Leibler divergence between the original and embedded distributions (van der Maaten and Hinton, 2008). ESM-1b and ESM-2-generated *T. b. brucei*

embeddings were concatenated with embeddings from one of the reference sets (generated from the same ESM model) and processed using scikit-learn's (v1.6.1) implementation of t-SNE. Models were run at perplexities 30 and 90 using an identical seed (random_state=42), while all other parameters were kept as default.

To evaluate the extent to which neighbour relationships are retained in the reduced dimensional space, a trustworthiness score was computed for each t-SNE representation. Trustworthiness measures the extent to which the structure of high-dimensional data is retained in a low-dimensional embedding (Venna and Kaski, 2001). It is calculated as:

$$T(k) = 1 - \frac{2}{nk(2n - 3k - 1)} \sum_{i=1}^{n} \sum_{j \in \mathcal{N}_i^k} \max\big(0, (r(i,j) - k)\big)$$

Where for each sample i, $\mathcal{N}_i^k$ is the set of k nearest neighbors of sample i in the output space, and $r(i,j)$ is the position of sample j among the ranked list of neighbours of i in the input space. In this study, we rank neighbour lists based on Euclidean distance. Trustworthiness is bound between 0 and 1, penalising top-k neighbours in output space that are not top-k neighbours in input space. We note an absence of guiding literature for what value constitutes a 'trustworthy' embedding, however, a preprint from Bombina et al. (2025) cites any value above 0.8 to be good.

2.2.3 Functional annotation with eggNOG-mapper:

To address the difficulty of evaluating t-SNE representations without ground truth labels, we used eggNOG-mapper v2.1.12 to infer enzyme functions via orthology-based transfer. Leveraging precomputed orthologous groups from the eggNOG v5 database, eggNOG-mapper assigned a variety of functional classifications to *T. b. brucei* proteins, including Gene Ontology terms, KEGG pathways, COG categories, and Enzyme Commission (EC) numbers (Cantalapiedra et al., 2021). To reduce the risk of false functional inference from distant orthologs, the taxonomic scope was

restricted to the Kinetoplastida class. Other settings were kept as default. See Box 1 for the full eggNOG-mapper query.

```
emapper.py --cpu 20 --mp_start_method forkserver --data_dir
/dev/shm/ -o out --output_dir
/emapper_web_jobs/emapper_jobs/user_data/MM_uoplf1v6 --temp_dir
/emapper_web_jobs/emapper_jobs/user_data/MM_uoplf1v6 --override -m
diamond --dmnd_ignore_warnings -i
/emapper_web_jobs/emapper_jobs/user_data/MM_uoplf1v6/queries.fasta
--evalue 0.001 --score 60 --pident 40 --query_cover 20 --
subject_cover 20 --itype proteins --tax_scope 5653 --
target_orthologs all --go_evidence non-electronic --pfam_realign
none --report_orthologs --decorate_gff yes --excel >
/emapper_web_jobs/emapper_jobs/user_data/MM_uoplf1v6/emapper.out 2>
/emapper_web_jobs/emapper_jobs/user_data/MM_uoplf1v6/emapper.err
```

Box 1: Command-line invocation of eggNOG-mapper used to assign functional annotations to *T. brucei* proteins. The query file (queries.fasta) was searched against the eggNOG database using DIAMOND with strict alignment thresholds (E-value ≤ 0.001, bit score ≥ 60, ≥40% sequence identity, ≥20% coverage) and restricted to the Kinetoplastida taxonomic group (NCBI taxon ID: 5653). All ortholog types were retrieved, Gene Ontology terms were filtered for non-electronic evidence, and results were exported in Excel format for downstream analysis.

2.2.4 Evaluation of Unsupervised learning methods:

Functional annotations obtained by eggNOG-mapper were used to further evaluate t-SNE and ESM model performance. Given that EC numbers classify enzymes by catalytic function, the presence of clusters homogenous for a specific EC number was viewed as an indicator of good model performance. KEGG pathways and COG categories, although also retrieved from eggNOG-mapper, were deemed unsuitable for this task as proteins within the same pathway can possess vastly different catalytic functions, while EC numbers were preferred over GO terms due to their hierarchical nomenclature. Visual assessment of EC clusters indicated low coverage; to identify homogenous clusters where EC numbers were absent, an interactive t-SNE plot was designed using Plotly Dash v3.1.1 to report the most common words in the FASTA descriptions of proteins within a user-defined boundary. Visual inspection of both automatic and manually annotated t-SNE plots allowed semi-quantitative assessment of how well functional information was captured.

2.2.5 Candidate selection:

Following evaluation of unsupervised learning steps, candidate methyltransferases were selected for molecular docking analysis manually from interactive t-SNE plots generated in Plotly v6.2.0 at perplexity 30, for both ESM-1b and ESM-2 embeddings, based on their proximity to tier 1 and 3 reference enzyme clusters. UniProt or NCBI IDs were matched to the *T. b. brucei* FASTA dataset to retrieve the corresponding amino acid sequences prior to molecular docking.

2.3 Structure-based molecular docking:

2.3.1 Co-folding with Boltz-2:

To verify the competency of the selected candidates for SAM-binding and methylation of the histone H3 tail, structural information is required. To achieve this, we used Boltz-2, an unsupervised framework that predicts multimolecular 3D structures based on multiple sequence alignments (MSAs). It encodes structural and interaction knowledge in an energy-based representation trained on both static and dynamic data and predicts binding affinity from the same internal representation it learns while modelling co-folding (Wohlwend et al., 2024; Passaro et al., 2025).

We performed co-folding under two different configurations: 1) the candidate protein with SAM, and 2) the candidate protein with SAM and histone H3. To facilitate consistent reference to these configurations in figures, figure captions, and results discussions, we applied the respective shorthand terms **C + S**, and **C + S + H3**.

To model C + S, we input both the candidate sequence and the SMILES string corresponding to SAM (CHEBI:15414). To model C + S + H3, these were additionally co-folded with the histone H3 amino acid sequence (UniProt ID: Q4GYX7). We note that the leading methionine residue in histone H3 is cleaved by *T. b. brucei* following translation and was therefore removed from the input sequence. MSAs were auto-generated using the mmseqs2 server by specifying the --use_msa_server argument

in the Boltz v2.2.0 command line implementation, while other parameters were left as default. Each run output a CIF file containing the 3D structure, and two JSON files containing 1) binding affinity and probability predictions, and 2) structural confidence scores. Confidence scores include template modelling (TM-scores), a measure of global similarity between a structural prediction and its hypothetical 'true' structure, and pLDDTs, a measure of confidence in each residue's local structure, averaged over all residues (Gomez and Kovalevskiy, 2024). Boltz-2 also outputs ipTM and ipLDDT scores, which measure only at protein–protein or protein–ligand interfaces.

2.3.2 Binding affinity prediction with GNINA:

To obtain another measure of ligand-binding affinity, candidate structures generated in section 2.3.1 were evaluated using GNINA v1.3.1. GNINA is a docking software that samples ligand conformations via Markov chain Monte Carlo (MCMC) chains and predicts their binding affinities using both the AutoDock Vina scoring function and convolutional neural network (CNN) models (Trott and Olson, 2009; McNutt et al., 2025). Structural pre-processing was performed as follows:

1. Co-folded complex structures (CIF format) were converted to PDB using Open Babel v3.1.1.

2. For C + S, PDB files were split into separate ligand and protein chains using pdbutils 2024.0.2.

3. For C + S + H3, the histone H3 and candidate protein chains were retained together, while the SAM ligand was isolated in a separate file.

For initial "in-place" scoring, GNINA was run with default parameters and the --score_only flag, preserving the original Boltz-2 poses. Because Open Babel alters protonation states during PDB conversion, ligands were subsequently reprocessed for re-docking as follows:

1. A representative ligand PDB file was manually corrected for protonation states in PyMOL v3.1.6.1, saved, and converted to PDBQT format using pdbutils.

2. The corrected ligand PDBQT was supplied via the -l argument, with the --autobox-ligand parameter set to the original ligand PDB (in the Boltz-2-predicted pose).

3. Docking was restricted to a single output pose (--num_modes 1).

GNINA outputs included the Vina binding affinities from the AutoDock Vina scoring function, CNN pose scores, and CNN binding affinities. While Vina scores predict the Gibbs free energy of binding (∆G) in kcal/mol, Boltz-2 affinities are given in log(IC_{50} µM) units (Wagen and Wagen, 2025). Thus, to allow comparison between the two methods, Boltz-2 affinities were converted into ∆G (kcal/mol) as follows:

$$\Delta G = -1.364 \times (6 - affinity)$$

2.4 Candidate filtering and ranking:

After comparing predicted Vina and Boltz-2 affinities, as well as Vina scores under different binding conditions, we chose re-docking Vina scores in the C+S configuration as the preferred metric on which to predict SAM-binders. Based on work by Sankar et al. (2023) reporting an average Vina score of -7.5 kcal/mol for known SAM-binders, we adopted a permissive threshold of -6.0 kcal/mol to include putative SAM-binders in the ranking set.

To distinguish general SAM-binders from those capable of methylating histone H3, we calculated the spatial distance between the SAM methyl donor group (sulphur-bound methyl) and the nearest nucleophilic nitrogen atom within the histone H3 sequence. This included the Ser-1 nitrogen (although not technically a side chain). Distance calculations were performed by merging the re-docked ligand pose (GNINA output) with the co-folded candidate and histone H3 complex (generated in section 2.3.1) and

using MDTraj v1.10.3 to a) identify the nearest nucleophilic nitrogen, b) identify the amino acid residue to which it belongs, and c) compute its distance to the SAM methyl group. Based on work by Liscombe, Louie, and Noel (2012), candidates positioning the SAM methyl group within 1–5 Å of a nucleophilic nitrogen were predicted as putative histone methyltransferases. We note that these thresholds should not be interpreted as definitive, but rather tools to guide candidate ranking and prioritisation.

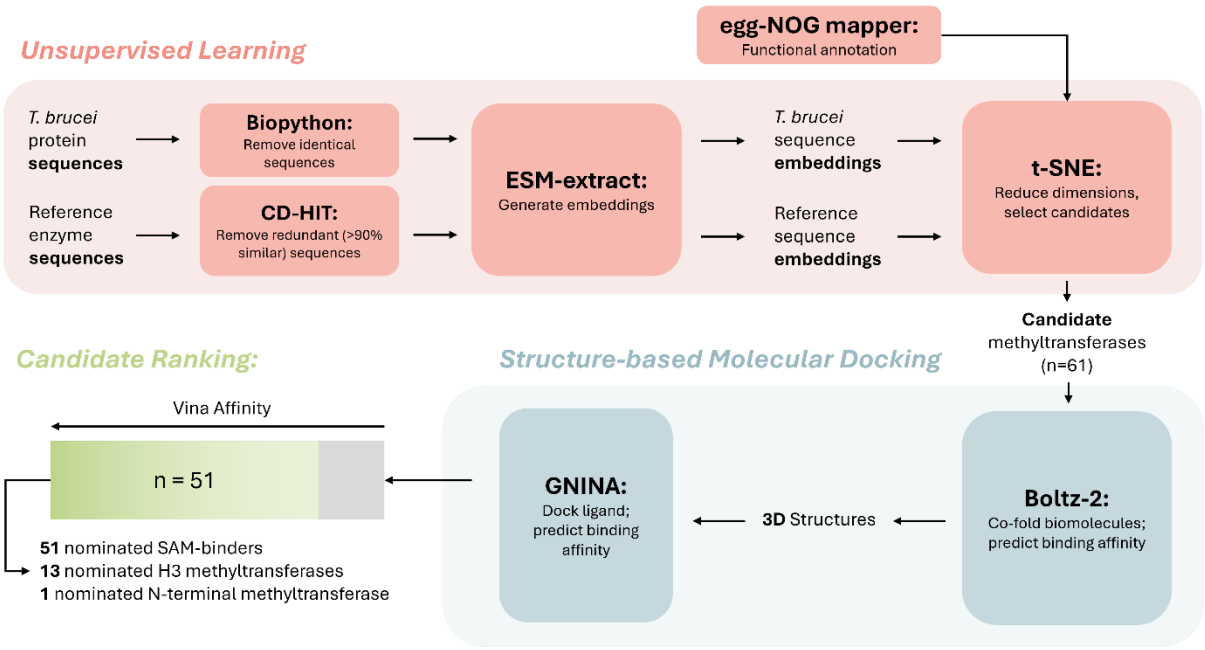


Figure 1: Graphical summary of study methodology.

3. Results:

3.1 Evaluation of t-SNE performance:

Prior to t-SNE visualisation, we scored each representation by trustworthiness to evaluate the extent to which the structure of the high-dimensional ESM embeddings is retained in the reduced dimensional space. Considering trustworthiness values of above 0.80 to be good (see section 2.2.2), Figure 2 shows that all representations remain trustworthy from $k = 5$ to $k = 50$, though as the number of nearest neighbours k considered by the scoring function increases, trustworthiness decreases for all representations. Above all other factors, t-SNE outputs more trustworthy projections for ESM2 than ESM1b. Among representations generated from the same ESM model, the relative trustworthiness of models run at perplexity 30 and 90 depends on the value of k. At $k = 5$ and $k = 10$, trustworthiness is higher at perplexity 30, before converging between $k = 15$ and $k = 25$. At $k \geq 25$, trustworthiness is highest for representations at perplexity 90. Lastly, neighbour relationships appear fractionally better preserved in tier 3 t-SNEs than tier 1 t-SNEs. These results suggest that all representations can be considered trustworthy within these neighbourhood sizes, though ESM-2 generated embeddings lead to more reliable 2D projections of protein similarity than ESM-1b, improving its interpretability for downstream tasks such as candidate selection and functional inference. The dependency of trustworthiness on k and perplexity indicate that candidate selection in t-SNE space is best contained within local neighbourhoods and at low perplexities.

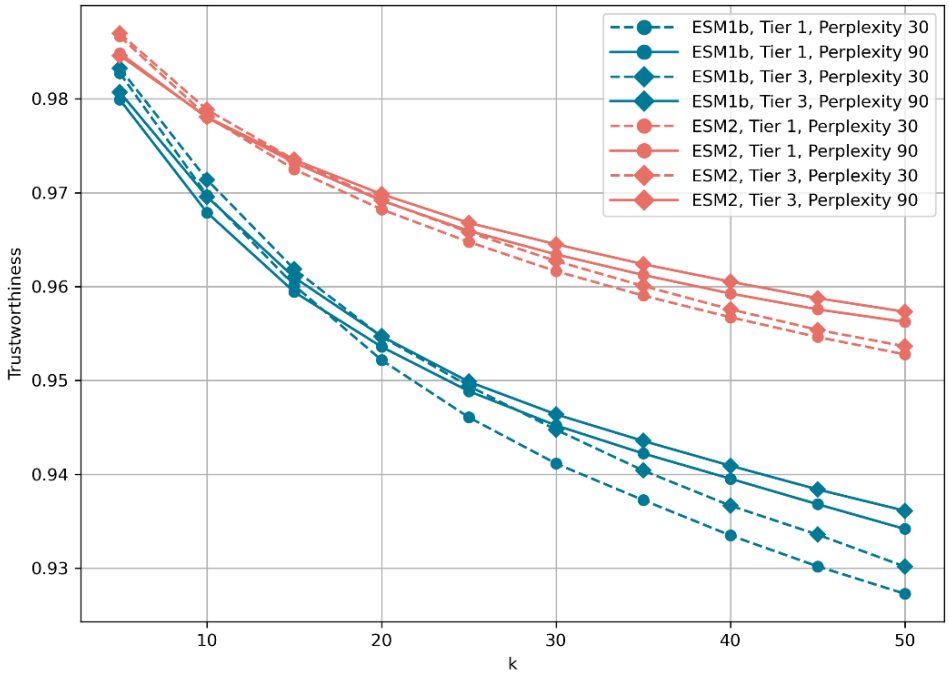


Figure 2: Trustworthiness of t-SNE representations. Trustworthiness was computed for t-SNE representations of ESM-1b and ESM-2 embeddings, at perplexities 30 and 90, for datasets containing tier 1 and 3 reference enzymes (together with the *T. b. brucei* dataset). k refers to the number of nearest neighbours considered by the trustworthiness scoring algorithm.

3.2 Evaluation of ESM and t-SNE with respect to biological significance:

In the absence of ground truth labels, assessing the extent to which ESM models capture biological information is difficult, and few conclusions can be drawn from intrinsic evaluation metrics. Therefore, *T. b. brucei* proteins were assigned Enzyme Commission (EC) numbers based on orthology using eggNOG-mapper. EC numbers group enzymes by catalytic activity in a hierarchical classification scheme. We therefore contend that ESM embeddings should visibly capture EC groups as well as reference methyltransferases in t-SNE representations. Figure 3 plots the *T. b. brucei* and reference protein embeddings after dimensionality reduction with t-SNE. In all representations (Figures 3A-D), while EC groups are not globally separable in 2D space, there are identifiable clusters corresponding to transferases (EC 2.) and lyases (EC 4.) as well as tier 1 and 3 reference enzymes. Small clusters comprising other EC groups are also visible, though these are frequently distributed among regions of high noise. Interestingly, no proteins were annotated as translocases (EC 7.) by eggNOG-

mapper. EC-labelled proteins appear to be concentrated on either the left or right of the plot, depending on the ESM model, though caution must be taken when referring to the dataset's global structure. It should be emphasized that some level of noise and lack of separation is expected; top-level EC numbers classify the type of catalysed reaction, such that each group still captures a diverse spectrum of substrates and protein structures. Therefore, the appearance of several clusters homogenous for a single EC group suggests that ESM does capture biologically relevant differences between enzyme families.

Since EC numbers cannot annotate the full complement of *T. b. brucei* proteins—such as those that play structural, signalling, regulatory, or transport functions—manual annotation of unlabelled proteins was performed with an interactive tool that reports the most common words in the FASTA descriptions of proteins within a user-defined boundary. This enabled exploratory analysis of biological significance. Figure 4A characterises several clusters of functionally distinct ESM-1b-embedded proteins, including adenylate cyclases (n = 64), mitochondrial carrier proteins (n = 29), amino acid transporters (n = 36), peptidyl-prolyl isomerases (n = 23), protein serine/threonine phosphatases (n = 17), kinases (n = 48), hexosyltransferases (n = 26), leucine-rich repeat proteins (n = 66) and ubiquitin-conjugating enzymes (n = 16). The ESM-2-embedded t-SNE (Figure 4B) identifies many of the same of functional groups and similar cluster sizes. Manual annotation of clusters in the centres of these t-SNE representations was less meaningful due to the lack of cluster distinction and high diversity of represented words.

The analysis also reveals clusters of proteins unique to Trypanosome biology. Clusters of retrotransposon hot spot (RHS) proteins (n = 40) and expression site-associated gene (ESAG) proteins appears in both Figure 4A and 4B, as well as three clusters of variable surface glycoproteins (VSGs) totalling 237 sequences in Figure 4A and 239 sequences in Figure 4B. Lastly, a large peripheral cluster appears in both ESM-1b and ESM-2-embedded plots, comprising approximately 900 proteins labelled as predominantly uncharacterised or *T. brucei*-specific. The ability of ESM embeddings to separate RHS, VSG, and other *T. brucei*-specific proteins suggests that ESM may capture subtle structural or evolutionary signals unique to kinetoplastids.

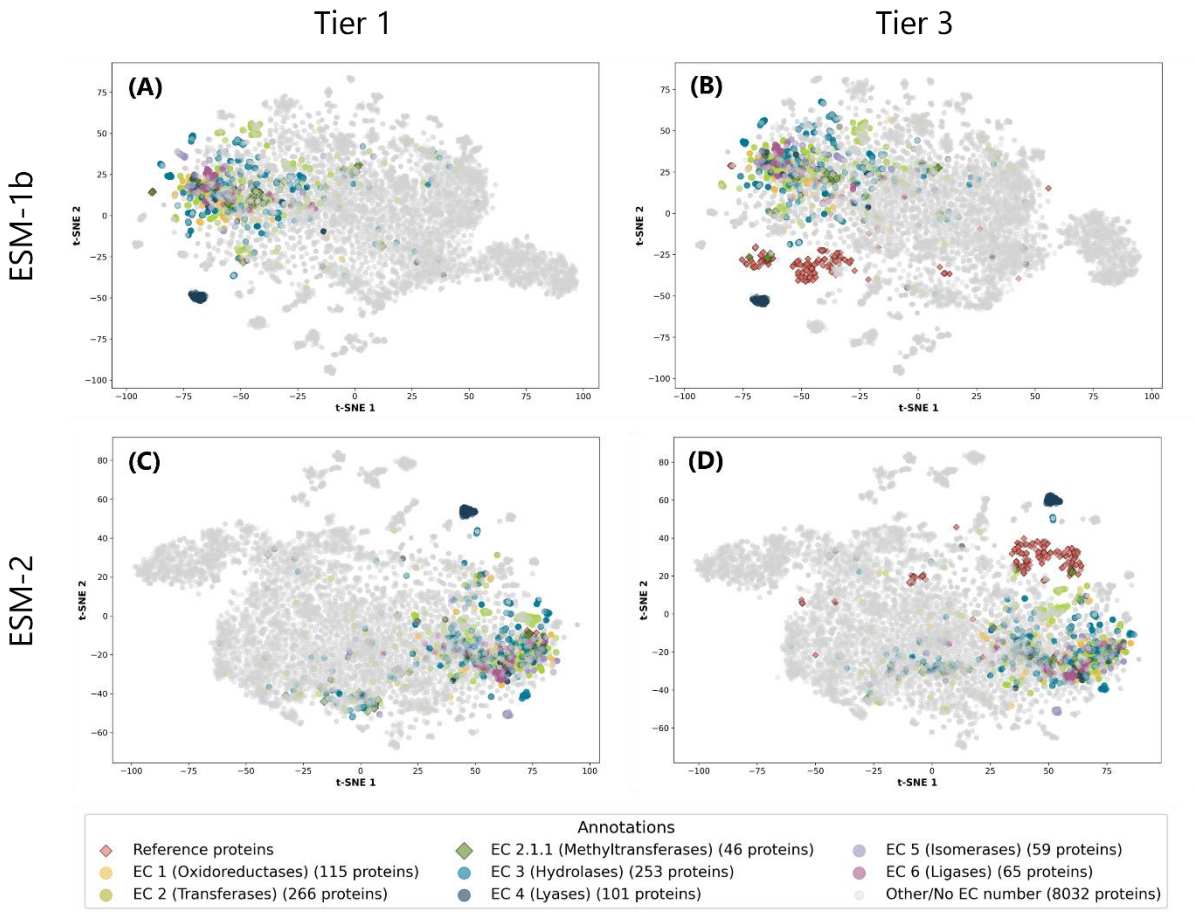


Figure 3: t-SNE projections of ESM embeddings (perplexity = 30). Projections differ by the choice of ESM model and reference enzyme dataset: **(A)** plots ESM-1b embeddings and reference N-terminal N-methyltransferases (tier 1); **(B)** plots ESM-1b embeddings and reference histone methyltransferases (tier 3); **(C)** plots ESM-2 embeddings and reference N-terminal N-methyltransferases (tier 1); **(D)** plots ESM-2 embeddings and refence histone methyltransferases (tier 3). Each data point represents a protein. Red diamonds indicate reference enzymes, while other colours distinguish Enzyme Commission (EC) numbers of *T. brucei* proteins. Proteins for which an EC number could not be retrieved are coloured grey.

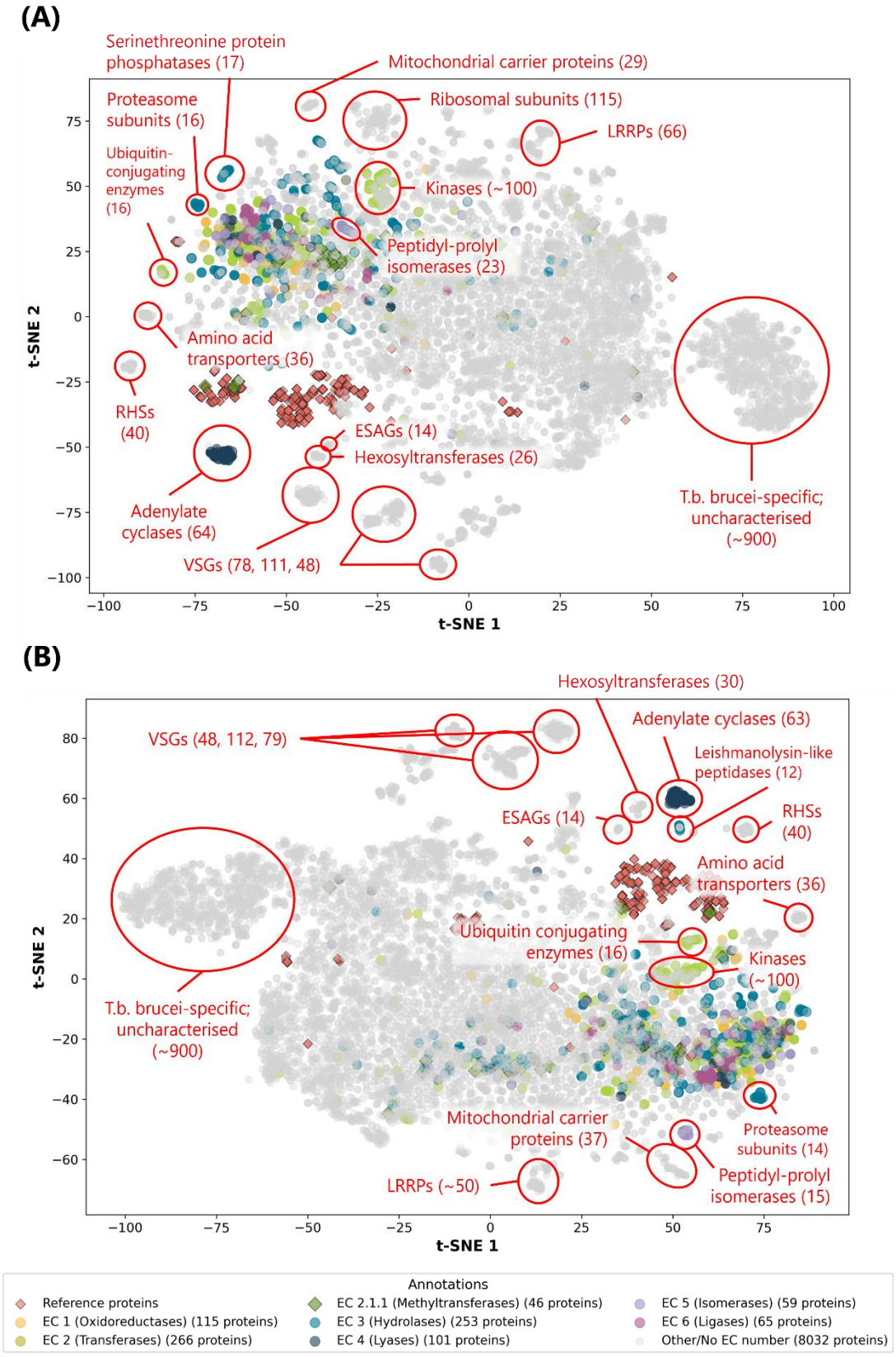


Figure 4: t-SNE projections of ESM-embeddings with manual annotations. (A) ESM-1b sequence embeddings with *T. brucei* and tier 3 reference datasets. **(B)** ESM-2 sequence embeddings with *T. brucei* and tier 3 reference datasets. Bracketed numbers denote cluster size.

3.3 Candidate selection:

Selection of candidates for molecular docking analysis was performed for all t-SNE representations in Figure 3. This was a manual process based on the proximity of *T. b. brucei* proteins to relevant clusters of reference enzymes in the plot. Figure 3 depicts the search spaces for each plot and highlights the selected candidates. Using ESM-1b, a single *T. b. brucei* protein (Q584S3) clustered with reference N-terminal N-methyltransferases (Figure 5A). With ESM-2, the same protein is the closest *T. brucei* protein to any N-terminal N-methyltransferase (Figure 5B). Figures 5B and 5D show candidates selected from clusters of histone methyltransferases, which yielded a larger candidate pool for downstream docking analysis. Importantly, there was disagreement between the ESM models regarding which proteins were most appropriate for selection; some *T. b. brucei* proteins clustered with reference enzymes embedded using one ESM model, but not the other, leading to some selected candidates appearing distant from reference clusters. In total, 61 proteins were selected for molecular docking analysis.

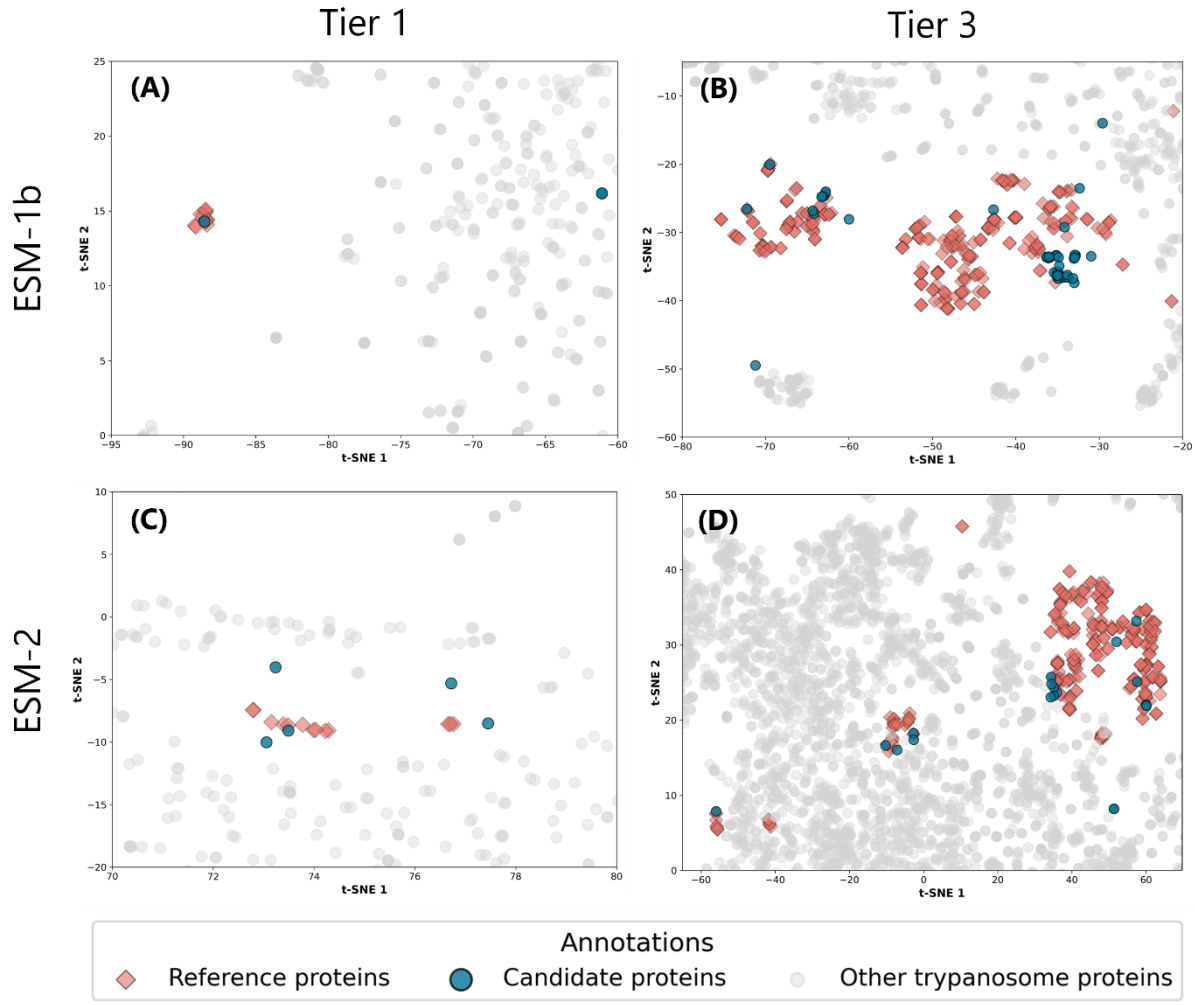


Figure 5: Candidate selection from t-SNE projections. Projections differ by the choice of ESM model and reference enzyme dataset: **(A)** plots ESM-1b embeddings and reference N-terminal N-methyltransferases (tier 1); **(B)** plots ESM-1b embeddings and reference histone methyltransferases (tier 3); **(C)** plots ESM-2 embeddings and reference N-terminal N-methyltransferases (tier 1); **(D)** plots ESM-2 embeddings and refence histone methyltransferases (tier 3). Each data point represents a protein. Red diamonds indicate reference enzymes, while blue circles denote selected candidates. Proteins not selected for further analysis are coloured grey.

3.4 Comparison of Boltz-2 and GNINA affinity predictions:

As Boltz-2 is a newly released co-folding software, its use invites comparison against established docking engines. AutoDock Vina, one of the most widely used molecular docking frameworks, was therefore selected as a benchmark. Figure 6 compares the

Gibbs free energy (ΔG) binding affinities predicted by Boltz-2 and GNINA (via its AutoDock Vina scoring function) of SAM with the candidate in the Boltz-2-predicted structure (C + S configuration). Overall, the results report a weak positive correlation (Pearson r = 0.503) between Boltz-2 and Vina predictions. The two methods agree most where Vina predicts moderate to strong binding, but where it predicts weak binding, Boltz-2 predictions remain moderate to strong. Consistent with this observation, the range of Boltz-2 predictions is smaller (-10.10 to -5.40; median = -7.20) than Vina's (-10.35 to -0.22; median = -6.66). When factoring in the binding probability value reported by Boltz-2, however, the results show that the level of disagreement is conditional on Boltz-2 binding probability. Indeed, Vina affinity and Boltz-2 probability present a strong correlation, with more energetically favourable candidate-ligand complexes exhibiting a higher binding probability (Pearson r = -0.784). Together, these observations indicate that Boltz-2 affinities are most reliable when interpreted in the context of their associated probabilities. Given its established status in the field, Vina affinity was adopted as the primary metric for subsequent candidate filtering.

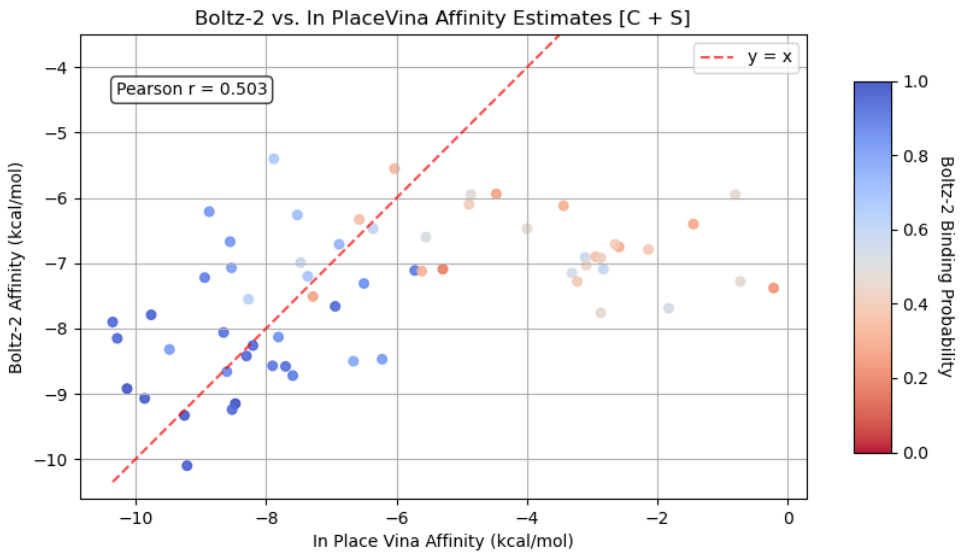


Figure 6: Comparison of in-place Vina and Boltz-2 binding affinity estimates. Values refer to the predicted binding affinity, in kcal/mol, of the ligand with the candidate protein in the Boltz-2-predicted 3D structure. Data point colours correspond to the binding probability as predicted by Boltz-2.

3.5 Binding conditions:

To determine the model conditions under which to predict candidates, Figure 7 visualises the impact of re-docking (Figure 7A) and modelling histone H3 (Figure 7B) on Vina binding affinities. Firstly, the results report a strong correlation (Pearson r = 0.830) between in place and re-docking affinities (C + S configuration). Using the line of y = x as reference, re-docking improves binding affinity compared to ligands scored in place for all but 6 proteins. These exceptions may arise from the fact that re-docking models were supplied with corrected SAM protonation states which weren't available for in place scoring (see section 2.3.2). Secondly, Figure 7B reports a strong correlation (Pearson r = 0.769) between re-docking affinities modelled with and without histone H3. Unlike in Figure 7A, the effect of modelling H3 does not appear to have a directionality (i.e it does not uniformly increase or decrease binding affinity). Given that re-docking both mitigates the risk of underestimating binding affinity from potentially inaccurate Boltz-2 predicted ligand orientations and models an improved SAM ligand, we choose to predict candidates under re-docked ligand conformations. Given the potential impact of modelling histone H3, even if small, we reserve histone H3 models for predicting histone H3 methyltransferases only.

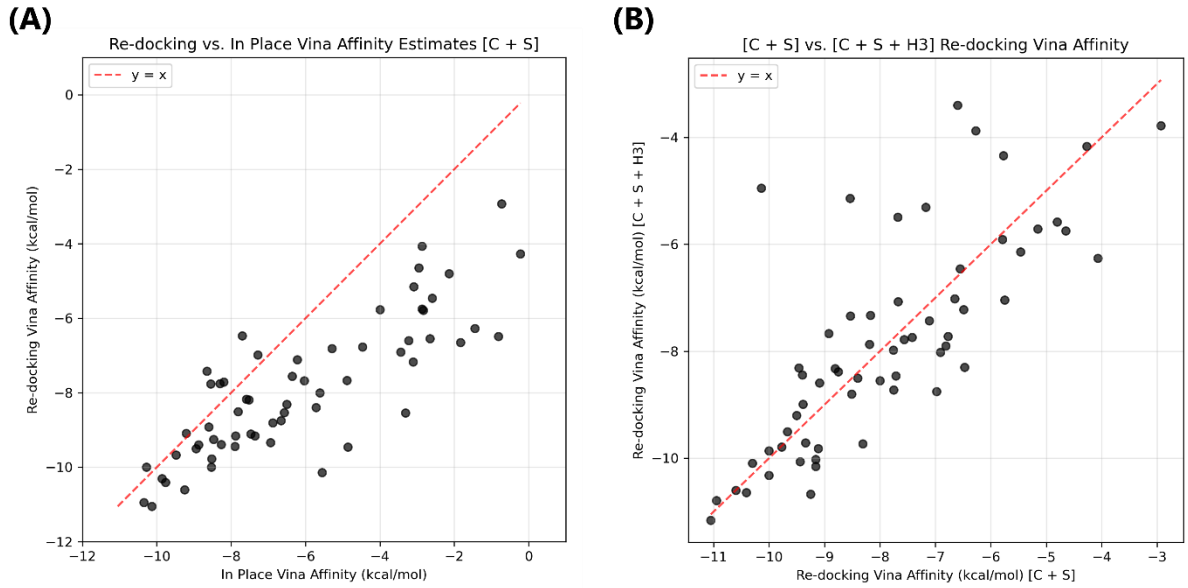


Figure 7: Comparison of GNINA docking conditions. (A) Comparison of in place and re-docking affinity predictions for C + S configuration. **(B)** Comparison of re-docking affinity predictions with (C + S + H3) and without (C + S) histone H3 modelling.

3.6 Predicting SAM-binders:

To generate a ranked list of proposed SAM-binders, we filtered all 61 candidates based on a Vina affinity threshold of ≥ -6 under re-docking conditions (corresponding to moderate-to-strong binding in the absence of histone H3). This step yielded 51 proteins (Figure 8). For the full ranked list of proteins, please refer to Supplementary Table 1.

3.7 Predicting histone H3 methyltransferases:

For methylation to proceed, the acceptor's nucleophilic atom must be positioned closer to the donor's methyl group than any other nucleophile in the vicinity, typically with about 3 Å of separation (Liscombe, Louie and Noel, 2012). Thus, to distinguish between histone H3 methyltransferases and proteins that are simply SAM-binding, we computed the distance between the SAM methyl donor and the nearest nitrogen methyl acceptor in histone H3 among all 51 candidates obtained in the previous step. Allowing an error of ± 2 Å, 13 proteins co-fold with histone H3 such that at least one of these histone residues is close enough to the methyl donor for methylation (Figure 8). The ranked list of predicted H3 methyltransferases and their inferred methylation sites are given in Table 1.

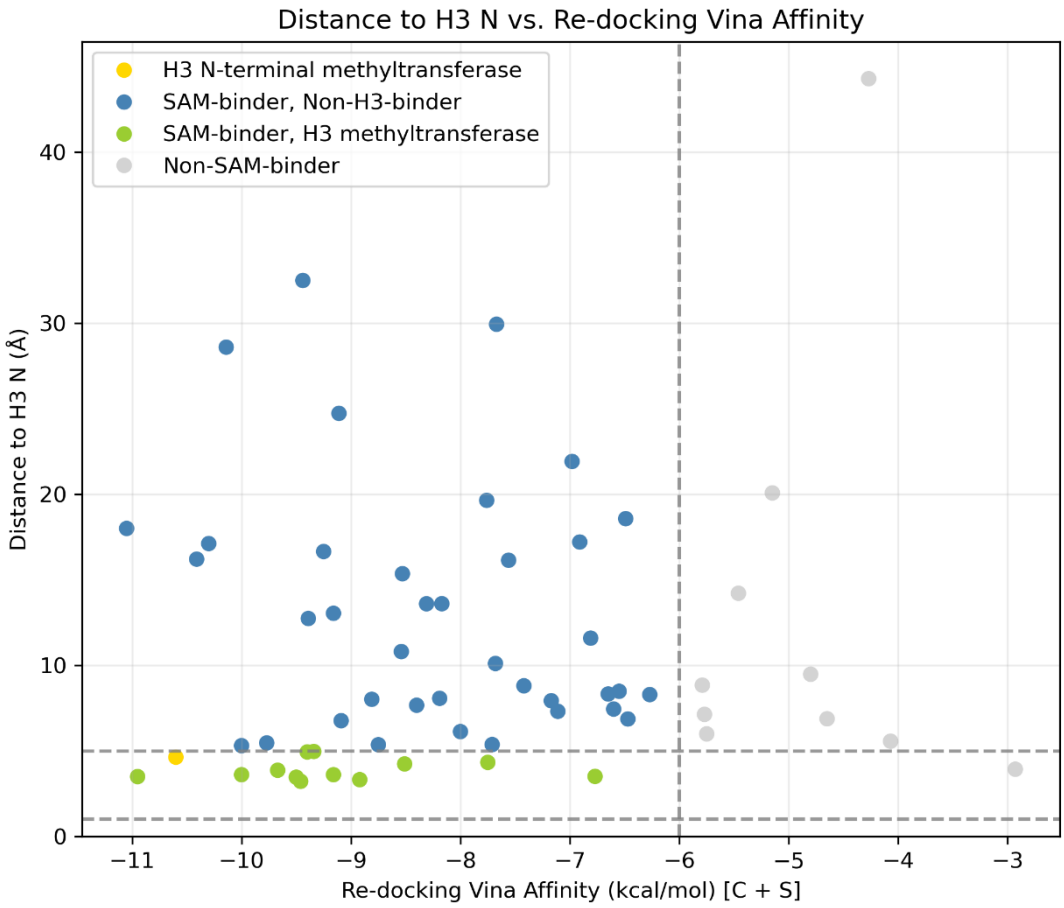


Figure 7: Prediction thresholds for SAM-binders and histone H3 methyltransferases. The vertical dashed line corresponds to the re-docking Vina affinity threshold (C + S configuration) of -7 kcal/mol around which proteins were predicted as SAM-binders (coloured) or non-SAM binders (gray). The horizontal dashed line corresponds to the distance window of 1-5 Å used to distinguish predicted histone H3 methyltransferases (green) with SAM-binders (blue) (after re-docking in the C + S + H3 configuration). Highlighted in gold is the protein nearest to the histone H3 Ser-1-N site.

Among proteins predicted as histone H3 methyltransferases (Table 1), re-docked Vina affinities (C + S + H3 configuration) range from -10.79 to -7.67 kcal/mol (median = -9.2), consistent with strong binding. Importantly, the most promising candidate for N-terminal methylation is Q584S3, with a distance of 4.6 Å between Ser-1-N and the SAM methyl donor. Other predicted methylation targets on the histone H3 tail include Arg-2-NH1, Lys-4-NZ, Arg-8-NH2, Lys-32-NZ, Arg-37-NH2, and Arg-39-NH2. Table 1 also reports the binding probabilities (median: 0.886) and ipTMs (median = 0.652) computed by Boltz-2 prior to re-docking. These do not translate to the re-docked structure and thus cannot be used for thresholding, serving instead as a

supplementary measure of confidence. While most candidates perform well, proteins Q581V5 and Q38C25 report binding probabilities of 0.472 and 0.443, respectively, and may represent false positives.

	Re-docked structure (C + S + H3)			Boltz-2 structure (C + S + H3)	
Protein ID	**Vina Score (kcal/mol)**	**Nearest Site**	**Distance (Å)**	**SAM Binding Probability**	**ipTM**
Q4GYA9	-10.79	ARG-2-NH1	3.5	0.953	0.615
Q584S3	-10.60	SER-1-N	4.6	0.977	0.661
Q38BP3	-10.15	ARG-2-NH1	3.6	0.815	0.535
Q57XC2	-9.86	LYS-4-NZ	3.6	0.886	0.642
Q582G4	-9.71	LYS-32-NZ	5.0	0.967	0.652
Q57XE0	-9.50	ARG-39-NH2	3.9	0.916	0.746
Q38AF4	-9.20	LYS-4-NZ	3.5	0.895	0.803
Q38A07	-8.80	ARG-39-NH1	4.2	0.907	0.791
Q57TW0	-8.72	LYS-4-NZ	4.3	0.831	0.669
pdb|6DNZ|D	-8.44	ARG-8-NH2	4.9	0.826	0.511
Q581V5	-8.31	LYS-4-NZ	3.2	0.472	0.649
Q38C25	-7.72	HIS-110-NE2	3.5	0.443	0.409
Q57UB6	-7.67	ARG-37-NH2	3.3	0.873	0.715

Table 1: Candidate methyltransferases for histone H3 residues, ranked by re-docking Vina affinity (kcal/mol). Protein ID refers to the UniProt ID or NCBI ID. Vina affinity corresponds to the estimated Gibbs free energy change (ΔG) in kcal/mol when SAM is re-docked into the candidate + H3 complex. The table also reports the identity of the nearest nitrogen nucleophile to the SAM methyl donor and the corresponding distance in angstrom (Å). Binding probability estimates the probability that the ligand binds the candidate protein, while ipTM measures the accuracy of the predicted relative positions of the subunits forming the protein-protein and protein-ligand interfaces (Wagen and Wagen, 2025). These metrics are both computed from the Boltz-2 predicted structure (prior to re-docking).

The formation of structurally robust intermolecular contacts is a necessary condition for binding (Ruiz-Carmona et al., 2016). However, thermodynamic stability is not a proxy for structural stability (Majewski, Ruiz-Carmona and Barril, 2019). Thus, we further examined the 3D structure of candidate–ligand–histone complexes for the top 6 ranked candidates in Table 1. The results identify several polar contacts, such as hydrogen bonds, between SAM and candidate chains. Figures 8A-E identify at least one interaction with the histone H3 acceptor site, while all candidates form several contacts with the ligand. Furthermore, no barriers to methylation are visible. Taken together, the pattern and density of polar contacts suggest the formation of stable and catalytically competent ternary complexes for each of these candidates.

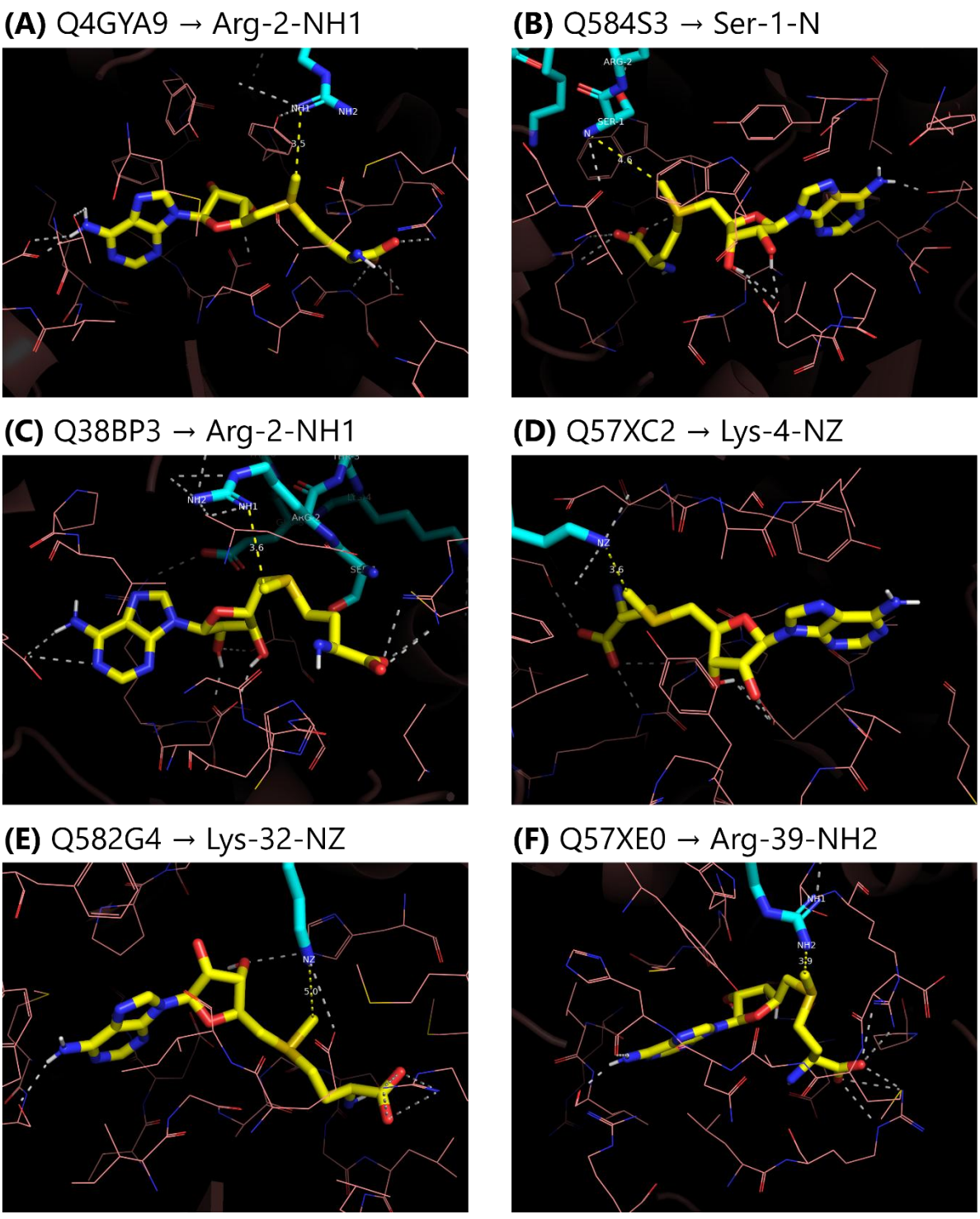


Figure 8: Visualisation of candidate–ligand–histone interactions. PyMOL visualisation of the best performing candidate methyltransferases by re-docked Vina affinity (C + S + H3). Yellow, cyan, and pink lines/sticks represent carbons atoms in SAM, histone H3, and the candidate protein, respectively. Blue, red and gold lines/sticks represent nitrogen, oxygen, and sulphur atoms, respectively. Yellow dashed lines show the distance (in Å) between the methyl donor and methyl acceptor. Grey dashed lines indicate polar contacts between chains.

4. Discussion:

4.1 On Unsupervised learning:

Quantitative evaluation of t-SNE projections using trustworthiness highlights established strengths and limitations of unsupervised dimensionality reduction. Figure 2 largely affirms the expected behaviour of t-SNE. For example, the notable decrease in trustworthiness with increasing neighbourhood size reflects t-SNE's inherent prioritisation of local structure over global relationships. The observation that trustworthiness is sensitive to the value of perplexity supports the consensus that perplexity models the trade-off between local and global structure—such that low perplexities better preserve local structure, and high perplexities better preserve global structure. Lastly, tier 3 datasets were more trustworthy than tier 1, likely due to the tier 3 dataset being larger and comprised of very similar proteins/embeddings whose relationships are retained in the t-SNE. Although using trustworthiness to evaluate the retention of local structure (i.e. up to a maximum neighbourhood size of 50) is sufficient for the purposes of candidate selection, complementing trustworthiness with global structure metrics such as Spearman rank correlation would provide a fuller picture of how well global ordering is retained. Overall, Figure 1 underlines that t-SNE is well suited for discovering functionally coherent local clusters, especially at lower perplexities. With the purpose of this study being to identify highly similar proteins, the choice to select candidates for molecular docking based on t-SNE representations at perplexity = 30 is justified.

Quantitative evaluation of t-SNE projections also lent insight into the underlying structure of ESM-1b and ESM-2 embeddings but underscored the challenge of using intrinsic evaluation metrics to extract biological relevance. ESM-2-derived t-SNE representations are more trustworthy than those generated from ESM-1b, regardless of the perplexity and dataset composition (Figure 2). This suggests that ESM-2 embeddings are more structured and locally coherent, and the level of distortion imposed on the dataset by the t-SNE algorithm is reduced. That this difference increases across a range of k implies that ESM2 improves medium- to larger-scale

relationships even more than local ones. As the newer model, this observation could reflect more biologically, functionally, or structurally meaningful embeddings. This interpretation is consistent with published benchmark data, where with the same number of parameters, ESM-2-predicted 3D structures achieve greater topological similarity between predicted and experimental protein structures than ESM-1b in CASP14 (TM-score: 0.51 vs. 0.42) and CAMEO (TM-score: 0.68 vs. 0.64) test sets (Lin et al., 2023). However, method-specific artefacts, such as a greater alignment between the ESM2 embeddings and t-SNE's projection bias, cannot be ruled out. The difficulty of evaluating ESM independently of t-SNE highlights ESM as a 'black-box' tool, where the meaning of each embedding dimension is abstract and not directly interpretable. Overall, we chose to select candidates based on both ESM-1b and ESM-2 embeddings to maximise the search space but consider ESM-2 to be the better option if a choice were necessary.

In the absence of ground truth labels, we sought to evaluate the ability of ESM and t-SNE to identify biological signal using EC annotations generated from eggNOG-mapper (Figure 3). The results indicated that a good level of biological signal emerges but also identify limitations of orthology-based annotation. We observed several clusters enriched for specific EC groups, though they mostly appear within regions of high noise. We contend that a significant source of noise is due to the labelling of proteins by top-level EC categories. Since EC numbers are hierarchical, top-level groups inflate intra-group heterogeneity, comprising enzymes with diverse substrates and mechanisms. This noise, therefore, does not necessarily indicate a lack of biological signal. Indeed, reference enzymes formed distinct and localised clusters, suggesting that their functional similarity is well captured. Orthology-based annotation has inherent limitations, however. While powerful, eggNOG-mapper is not a substitute for ground truth labels; outcomes are sensitive to the choice of taxonomic scope, e-value and identity thresholds, as well as the quality of database annotations. For divergent proteins, EC assignments may be misleading, as illustrated by the absence of proteins annotated as translocases (EC 7) by eggNOG-mapper (Mani et al., 2015). Complementing orthology with domain-level predictions and consensus annotation (e.g. InterProScan) may increase robustness, but ultimately, ESM is best benchmarked against ground truth datasets.

Orthology-based annotation was complemented by manual annotation using common FASTA terms, a strategy which further identified biological signal and alluded to ESM's potential as an exploratory tool. Through this method (Figure 4), we were able to identify peripheral clusters corresponding to individual protein families (e.g adenylate cyclases, mitochondrial carrier proteins, peptidyl-prolyl isomerases, ubiquitin-conjugating enzymes), functional groups (e.g kinases, amino acid transporters, hexosyltransferases), and structural motif–defined superfamilies (e.g leucine-rich repeat proteins), as well as Trypanosome-specific ESAG, RHS and VSG proteins (Graham and Barry, 1991; Florini et al., 2018; Cross, Kim and Wickstead, 2014). This further substantiates that ESM + t-SNE can differentiate between functionally, structurally, and evolutionarily distinct proteins. However, the subjective nature of manual cluster definition and concerns about FASTA header quality complicate quantitative claims and further argue for multi-modal validation.

Observations from manual annotation also raise wider questions about Trypanosome biology. The three distinct VSG clusters (together totalling 237-239 sequences) appearing in all t-SNE representations could reflect subtype stratification, different expression states, or sequence motifs shared across subsets of VSGs (Figure 4). A 2013 genomic survey (https://tritrypdb.org/common/downloads/release-5.0/) estimated ~1274 VSG-like sequences, of which ~1000 are pseudogenes and fewer than 200 are putative complete or near-complete VSGs (Cross, Kim and Wickstead, 2014). The large peripheral "peninsula" of *T. brucei*-specific/uncharacterised proteins might therefore represent translated VSG pseudogenes, though they could also be lineage-specific expansions or assembly artefacts. Finally, contrasting tight, separable "island clusters" with those on the "mainland" (which are much harder to annotate manually) could reveal whether islands reflect functional coherence, evolutionary distance, or projection artefacts. Follow-up analyses to explore these possibilities could include domain composition screening and phylogenetic profiling.

To summarise the unsupervised learning stage, trustworthiness analysis indicates that local neighbourhoods of similar ESM embeddings are well-preserved by t-SNE, especially at low perplexities, while orthology-based EC annotations and manual FASTA-based cluster labelling demonstrate that t-SNE representations of ESM embeddings capture biologically meaningful features. However, due in part to

previously discussed limitations of automated annotation, this signal is hard to quantify, prone to noise and not exhaustive. With this in mind, we propose alternative intrinsic scoring metrics and annotation strategies. Additionally, other dimensionality reduction methods such as UMAP and Neighbour Joining (NJ) trees could help reveal relationships between protein families that, due to distortion, are not interpretable with t-SNE. Both methods better preserve global structure; UMAP's loss function actively repels distant points, while NJ trees construct a hierarchical representation of relationships by iteratively joining pairs of sequences that minimise total branch length (Saitou and Nei, 1987). NJ trees based on ESM-1b embeddings have been successful in hierarchical protein classification problems, including for SAM-binding enzymes (Yeung et al., 2023).

4.2 On candidate selection:

Though they do not affect the findings from molecular docking analysis, the candidate selection process represents a key limitation of this study. Firstly, selecting docking candidates directly from t-SNE plots introduces avoidable technical biases (Figure 5). Although local relationships are well-preserved, any level of distortion imposed by t-SNE means that that prioritising proteins by 2D proximity risks increasing the number of false positives and omitting true neighbours. Secondly, selecting candidates manually and without quantitative measurement introduces an additional source of bias and limits reproducibility. It is important to note that because the co-folding and scoring of each candidate is independent, these limitations do not weaken the quality of molecular docking data or undermine the findings of this study. While they do hinder the application of this workflow as an automated and standardised pipeline, this can be overcome with alternative candidate selection strategies. For example, selecting nearest neighbours in the full ESM embedding space using reproducible, mathematical criteria (e.g. cosine similarity or Euclidean distance with a fixed radius, or k-NN with deterministic tie-breaking) and reporting thresholds (similarity, k) would make candidate selection transparent and repeatable. t-SNE has much value as a visualisation tool; it effectively conveys underlying data structures and biological insight. Thus, it should be used to illustrate selections that were made quantitatively in

embedding space rather than drive selection. With these changes, this unsupervised learning workflow could be applied at scale to identify novel proteins based on functional, structural, and evolutionary similarity, without the need for large structural datasets and multiple sequence alignment.

4.3 On structure-based molecular docking:

Comparison of Boltz-2 and GNINA (Vina) affinity predictions highlight complementary strengths but also underscore the need for benchmarked calibration of Boltz-2-predicted metrics. In our data (Figure 6), Boltz-2 compresses affinity ranges, agreeing with Vina only for moderate-to-strong binders, suggesting that Boltz-2 may under-represent weak binding due to its training distribution. That agreement improves when Boltz-2's numerical affinity output is interpreted in the context its binding probability metric suggests that the scalar Boltz-2 affinity should be interpreted conditional on a high probability score; when probability is low, we should de-weight the affinity. An important caveat when comparing numbers is that the two scales are not exactly commensurate: Boltz-2's scalar is a learned, IC50-like value on a log(μM) scale trained on many kinds of lab readouts, whereas Vina's score is a pose-dependent empirical estimate of ΔG in kcal/mol. Therefore, our conversion of Boltz-2 into kcal/mol is a readability bridge and not a measure of physical equivalence.

While this co-folding workflow already incorporates histone H3 to contextualise docking, it does not probe for other important biological variables. Some of these, such as catalytic by-products and DNA interactions, can be examined within our existing toolkit. For example, S-adenosyl-homocysteine—the by-product of SAM-dependent methylation—can competitively inhibit SAM or remain bound after methylation, suggesting it may exhibit distinct binding signatures (Liscombe, Louie and Noel, 2012). Testing binding of S-adenosyl-homocysteine (SAH) against SAM would thus help discriminate catalytic interactions and reveal whether active-site geometry and thermodynamics support catalysis. Secondly, Boltz-2 method conditioning allows structural predictions for DNA-protein complexes, important in the context of histone methylation. Boltz-2 also offers further user-controllability which could be leveraged to

produce more accurate predictions. For example, incorporating structural templates, pocket conditioning, and distance restraints may enforce more biologically plausible methyl-donor/acceptor geometry (Passaro et al., 2025). Other relevant biological factors require investigation with other methods. Boltz-2's static structure prediction does not account for flexible loops, allosteric shifts, or induced-fit binding. Conditioning predictions with short molecular dynamics (MD) simulations could help account for these effects, allowing refinement of binding poses and more realistic sampling of flexible active sites.

Overall, Boltz-2's functionality is well suited to protein identification tasks but trades the accuracy of gold standard physics-based methods for computational efficiency. As a co-folding software, there is impressive flexibility in the number and type of relevant biomolecules that can be modelled. In this study, incorporating histone H3 into structural predictions creates a fuller picture of *in vivo* conditions and enables the prediction of methylation targets. Being 1000x faster than gold-standard free energy perturbation (FEP) methods, Boltz-2 is also practical for high-throughput pipelines (Passaro et al., 2025), and its ability to produce integrated structure, hit probability, and affinity predictions could greatly streamline the screening process. Although powerful, Boltz-2's quantitative accuracy for novel targets remains under active evaluation. Like other deep learning models, its predictions can be biased towards the distribution of its training set, and systematic benchmarking across diverse protein–ligand classes is still limited in the literature. Early investigations place Boltz-2 as lagging behind FEP methods for affinity prediction, and comparable with AlphaFold3 for structure prediction, though performance varies by protein family and substrate (Passaro et al., 2025; Ille et al., 2025). A formal benchmark—using a curated dataset with known SAM binders/non-binders spanning a range of affinities—would clarify the sensitivity and specificity of Boltz-2's affinity and probability outputs.

4.4 Nominated methyltransferases and future work

A recurring theme in this work is the reliance on permissive or empirically derived thresholds for filtering candidates. While the values chosen are justified by the literature, thresholding as a strategy has inherent limitations. Firstly, Vina does not

provide a binary “binder or non-binder” classification; rather, it estimates the thermodynamic favourability of a given pose. As such, weak binders—falling above a permissive threshold—may still be biologically relevant, particularly in the context of transient interactions or binding that depends on cofactors absent from the model. Conversely, strong predicted binders may be false positives if the pose is non-physiological. While Boltz-2 does report binding probability as well as affinity values, guidance for their interpretation is lacking, and it remains unclear what constitutes “strong” or “weak” evidence in Boltz-2’s native units across diverse protein–ligand classes. Secondly, given the large number of metrics output by Boltz-2 and GNINA, it is difficult to efficiently unify them to guide functional predictions. Boltz-2 metrics such as pLDDT, ipLDDT, pTM, and ipTM, as well as GNINA CNN scores, remain largely unexplored in this study. At present, the values we report are best interpreted in relative rather than absolute terms, and only in combination, to guide and prioritise novel enzymes for experimental validation.

Examination of FASTA annotations in the ranked list of nominated methyltransferase (Table 1) revealed a strong enrichment for proteins with known or putative methyltransferase activity, supporting the biological plausibility of our pipeline. Firstly, 6 candidates are referred to as confirmed or putative methyltransferases, of which 5 are the highest ranked by Vina affinity. This includes the verified protein arginine N-methyltransferase 7 (TbPRMT7; UniProt ID: Q582G4), which has been shown to methylate a variety of substrates including histones (Fisk et al., 2009). The other methyltransferase annotations are unreviewed; thus, the results simultaneously lend credence to these annotations and our pipeline. However, predicted methylation targets occasionally disagree, as in the case of Q57XC2 (labelled as arginine methyltransferases, predicted target lysine-4), and TbPRMT7 (predicted target lysine-32). A further 6 candidates are annotated as SET domain–containing. The SET domain is approximately 130 amino acids long and strongly associated with lysine methylation in histones and other proteins (Rea et al., 2000; Yeates, 2002; Dillon et al., 2005). Most promisingly, our pipeline identified a putative alpha N-terminal methyltransferase which clustered with reference N-terminal methyltransferases in the embedded space, docked SAM favourably, and co-folded to orient SAM close to H3-Ser-1. Overall, the ranked list of nominated proteins favours a pipeline capable of yielding novel methyltransferases in *T. b. brucei*.

Following conclusion of this analysis, collaborators in the Akiyoshi Laboratory (University of Edinburgh) independently identified a protein (Q38BE3) as a potential histone H3 N-terminal methyltransferase using alternative computational approaches and experimental validation. Q38BE3 does not localise with either reference histone methyltransferases or reference N-terminal methyltransferases in the reduced dimensional space (Supplementary Figure 1), perhaps reflecting atypical sequence motifs that diverge from canonical methyltransferases but still enable similar function (convergent evolution), or the prioritisation of features by ESM/t-SNE that are not directly related to the catalytic mechanism. Though not oriented to methylate H3-Ser-1 by structural analysis (nearest residue: Gln-82; distance: 19.5 Å), indicating that a catalytically competent ternary complex was not formed, it did have a strong predicted binding affinity for SAM (-9.58 kcal/mol [C + S]). It is important to note that the experimental data pertaining to Q38BE3 is preliminary and awaits peer-review, but this observation emphasises that similarity in ESM embedding space may not be an absolute requirement for shared function.

A final limitation of this study is the assumption that the methyltransferase targets of interest are S-adenosyl methionine (SAM)–dependent. While SAM is the most common methyl group donor in biological systems, used by over 95% of all methyltransferases (Lennard and Wang, 2018), some instead utilise alternative cofactors such as S- methyltetrahydrofolate (MTHF) and methyl-B12 (Bare et al., 2023). This study therefore shoulders a small risk that the target N-terminal methyltransferase uses a non-canonical methyl donor, for which this pipeline is not appropriate.

Since our computational methods are hypothesis-generating rather than confirmative, future work should involve both methodological refinement and targeted experimental follow-up. Pipeline refinement should incorporate the improvements discussed above, such as quantitative, embedding-space-based candidate selection, complementary dimensionality reduction approaches, and structural modelling under different controls and interacting molecules. Experimental strategies should prioritise highly ranked proteins with predicted H3 methylation sites. Suitable approaches include targeted

mass spectrometry and genetic perturbation experiments (e.g., knockouts or knockdowns) to assess phenotypic consequences of candidate loss.

5. Conclusion:

In this report, we demonstrate that combining protein language model embeddings with non-linear dimensionality reduction and structure-based molecular docking provides a tractable pipeline for nominating candidate methyltransferases in *Trypanosoma brucei brucei*, while also revealing the methodological trade-offs that accompany unsupervised discovery in a largely unlabelled proteome and recommending improvements to enhance its robustness and scalability. Integrating orthology-based and manual annotation of embedded t-SNE representations revealed biologically coherent local clusters; evidence of the robust biological signal contained within ESM embedding profiles. The localisation of reference methyltransferase clusters provided a refined search space from which a shortlist of proteins was selected for molecular docking analysis. By independently assessing each candidate-histone-ligand complex for its binding affinity, probability, and structural stability, several plausible methyltransferases and their predicted targets emerged, including a strong candidate for N-terminal histone H3 methylation. Beyond trypanosomes, this study contributes a novel, scalable pipeline for enzyme discovery that can be readily adapted to other enzymatic functions, species, and molecular systems. As computational models and structural prediction tools continue to improve, approaches like this will become increasingly powerful engines for characterising life's machinery.

6. References:

Akiyoshi, B. and Gull, K. (2014). Discovery of Unconventional Kinetochores in Kinetoplastids. *Cell*, 156(6), pp.1247–1258. doi:https://doi.org/10.1016/j.cell.2014.01.049.

AlphaFold and beyond. (2023). *Nature Methods*, 20(2), pp.163–163. doi:https://doi.org/10.1038/s41592-023-01790-6.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3), pp.403–410. doi:https://doi.org/10.1016/S0022-2836(05)80360-2.

Anfinsen, C.B. (1973). Principles that Govern the Folding of Protein Chains. *Science*, [online] 181(4096), pp.223–230. doi:https://doi.org/10.1126/science.181.4096.223.

Avery, C., Patterson, J., Grear, T., Frater, T. and Jacobs, D.J. (2022). Protein Function Analysis through Machine Learning. *Biomolecules*, [online] 12(9), p.1246. doi:https://doi.org/10.3390/biom12091246.

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., Cong, Q., Kinch, L.N., Schaeffer, R.D., Millán, C., Park, H., Adams, C., Glassman, C.R., DeGiovanni, A., Pereira, J.H., Rodrigues, A.V., van Dijk, A.A., Ebrecht, A.C. and Opperman, D.J. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557), pp.871–876. doi:https://doi.org/10.1126/science.abj8754.

Bare, A., Thomas, J., Etoroma, D. and Lee, S.G., 2023. Functional analysis of phosphoethanolamine N-methyltransferase in plants and parasites: Essential S-adenosylmethionine-dependent methyltransferase in choline and phospholipid metabolism. In: J. Jez, ed. *Methods in Enzymology*. 1st ed. London: Academic Press, 680, pp.101–137. doi:10.1016/bs.mie.2022.08.028.

Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Adesina, A., Ahmad, S., Bowler-Barnett, E.H., Bye-A-Jee, H., Carpentier, D., Denny, P., Fan, J., Garmiri, P., Jose, L., Hussein, A., Ignatchenko, A., Insana, G., Ishtiaq, R., Joshi, V., Jyothi, D. and Kandasaamy, S. (2024). UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Research*, 53(D1). doi:https://doi.org/10.1093/nar/gkae1010.

Bombina, P., Adams, Z.B., McGee, R.L. and Coombes, K.R. (2025). From Pairwise Distances to Neighborhood Preservation: Benchmarking Dimensionality Reduction Algorithms for CyTOF, scRNA-seq, and CITE-seq. *bioRxiv (Cold Spring Harbor Laboratory)*. doi:https://doi.org/10.1101/2025.04.28.651069.

Brandes, N., Goldman, G., Wang, C.H., Ye, C.J. and Ntranos, V. (2023). Genome-wide prediction of disease variant effects with a deep protein language model. *Nature Genetics*, 55. doi:https://doi.org/10.1038/s41588-023-01465-0.

Cantalapiedra, C.P., Hernández-Plaza, A., Letunic, I., Bork, P. and Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution*, 38(12), pp.5825–5829. doi:https://doi.org/10.1093/molbev/msab293.

Chen, J.-Y., Wang, J.-F., Hu, Y., Li, X.-H., Qian, Y.-R. and Song, C.-L. (2025). Evaluating the advancements in protein language models for encoding strategies in protein function prediction: a comprehensive review. *Frontiers in Bioengineering and Biotechnology*, 13. doi:https://doi.org/10.3389/fbioe.2025.1506508.

Chen, L., DeVries, A.L. and Cheng, C.C. (1997). Convergent evolution of antifreeze glycoproteins in Antarctic notothenioid fish and Arctic cod. *Proceedings of the National Academy of Sciences*, 94(8), pp.3817–3822. doi:https://doi.org/10.1073/pnas.94.8.3817.

Cross, G.A.M., Kim, H.-S. and Wickstead, B. (2014). Capturing the variant surface glycoprotein repertoire (the VSGnome) of *Trypanosoma brucei* Lister 427. *Molecular and Biochemical Parasitology*, [online] 195(1), pp.59–73. doi:https://doi.org/10.1016/j.molbiopara.2014.06.004.

Deák, G., Wapenaar, H., Sandoval, G., Chen, R., Taylor, M.R.D., Burdett, H., Watson, J.A., Tuijtel, M.W., Webb, S. and Wilson, M.D. (2023). Histone divergence in trypanosomes results in unique alterations to nucleosome structure. *Nucleic Acids Research*, 51(15), pp.7882–7899. doi:https://doi.org/10.1093/nar/gkad577.

Diaz, K., Meng, Y. and Huang, R. (2021). Past, present, and perspectives of protein N-terminal methylation. *Current Opinion in Chemical Biology*, 63, pp.115–122. doi:https://doi.org/10.1016/j.cbpa.2021.02.017.

Dillon, S.C., Zhang, X., Trievel, R.C. and Cheng, X. (2005). The SET-domain protein superfamily: protein lysine methyltransferases. *Genome Biology*, 6(8), p.227. doi:https://doi.org/10.1186/gb-2005-6-8-227.

Fisk, J.C., Sayegh, J., Zurita-Lopez, C., Menon, S., Presnyak, V., Clarke, S.G. and Read, L.K. (2009). A Type III Protein Arginine Methyltransferase from the Protozoan Parasite *Trypanosoma brucei*. *Journal of Biological Chemistry*, 284(17), pp.11590–11600. doi:https://doi.org/10.1074/jbc.m807279200.

Florini, F., Naguleswaran, A., Gharib, W.H., Bringaud, F. and Roditi, I. (2018). Unexpected diversity in eukaryotic transcription revealed by the retrotransposon hotspot family of *Trypanosoma brucei*. *Nucleic Acids Research*, 47(4), pp.1725–1739. doi:https://doi.org/10.1093/nar/gky1255.

Forli, S., Huey, R., Pique, M.E., Sanner, M.F., Goodsell, D.S. and Olson, A.J. (2016). Computational protein–ligand docking and virtual drug screening with the AutoDock suite. *Nature Protocols*, 11(5), pp.905–919. doi:https://doi.org/10.1038/nprot.2016.051.

Gligorijević, V., Renfrew, P.D., Kosciolek, T., Leman, J.K., Berenberg, D. and Vatanen, T. (2021). Structure-based protein function prediction using graph convolutional networks. *Nature Communications*, 12(1). doi:https://doi.org/10.1038/s41467-021-23303-9.

Gomez, P.G.M. and Kovalevskiy, O. (2024). *AlphaFold: A practical guide*. [online] EMBL-EBI. Available at: https://doi.org/10.6019/tol.alphafold-w.2024.00001.1 [Accessed 13 Apr. 2025].

Graham, S.V. and Barry, J.D. (1991). Expression site-associated genes transcribed independently of variant surface glycoprotein genes in *Trypanosoma brucei*. *Molecular and Biochemical Parasitology*, 47(1), pp.31–41. doi:https://doi.org/10.1016/0166-6851(91)90145-v.

He, Y., Zhou, X., Chang, C., Chen, G., Liu, W., Li, G., Fan, X., Sun, M., Miao, C., Huang, Q., Ma, Y., Yuan, F. and Chang, X. (2024). Protein language models-assisted optimization of a uracil-N-glycosylase variant enables programmable T-to-G and T-to-C base editing. *Molecular Cell*, 84(7), pp.1257-1270.e6. doi:https://doi.org/10.1016/j.molcel.2024.01.021.

Hittinger, C.T. and Carroll, S.B. (2007). Gene duplication and the adaptive evolution of a classic genetic switch. *Nature*, 449(7163), pp.677–681. doi:https://doi.org/10.1038/nature06151.

Hu, M., Alkhairy, S., Lee, I., Pillich, R.T., Fong, D., Smith, K., Bachelder, R., Ideker, T. and Pratt, D. (2024). Evaluation of large language models for discovery of gene set function. *Nature Methods*, [online] 22. doi:https://doi.org/10.1038/s41592-024-02525-x.

Ille, A.M., Markosian, C., Burley, S.K., Pasqualini, R. and Arap, W. (2025). Human protein interactome structure prediction at scale with Boltz-2. *bioRxiv*. doi:https://doi.org/10.1101/2025.07.03.663068.

Ishii, M. and Akiyoshi, B. (2020). Characterization of unconventional kinetochore kinases KKT10 and KKT19 in *Trypanosoma brucei*. *Journal of Cell Science*, 133(8). doi:https://doi.org/10.1242/jcs.240978.

Jakobsson, M.E., Małecki, J.M., Halabelian, L., Nilges, B.S., Pinto, R., Kudithipudi, S., Munk, S., Davydova, E., Zuhairi, F.R., Arrowsmith, C.H., Jeltsch, A., Leidel, S.A.,

Olsen, J.V. and Falnes, P.Ø. (2018). The dual methyltransferase METTL13 targets N terminus and Lys55 of eEF1A and modulates codon-specific translation rates. *Nature Communications*, 9(1). doi:https://doi.org/10.1038/s41467-018-05646-y.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J. and Back, T. (2021). Highly Accurate Protein Structure Prediction with Alphafold. *Nature*, [online] 596(7873), pp.583–589. doi:https://doi.org/10.1038/s41586-021-03819-2.

Kraus, A.J., Vanselow, J.T., Lamer, S., Brink, B.G., Schlosser, A. and Siegel, T.N. (2020). Distinct roles for H4 and H2A.Z acetylation in RNA transcription in African trypanosomes. *Nature Communications*, 11(1). doi:https://doi.org/10.1038/s41467-020-15274-0.

Li, W. and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13), pp.1658–1659. doi:https://doi.org/10.1093/bioinformatics/btl158.

Lin, B., Luo, X., Liu, Y. and Jin, X. (2024). A comprehensive review and comparison of existing computational methods for protein function prediction. *Briefings in Bioinformatics*, 25(4). doi:https://doi.org/10.1093/bib/bbae289.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S. and Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637), pp.1123–1130. doi:https://doi.org/10.1126/science.ade2574.

Liscombe, D.K., Louie, G.V. and Noel, J.P. (2012). Architectures, mechanisms and molecular evolution of natural product methyltransferases. *Natural Product Reports*, 29(10), p.1238. doi:https://doi.org/10.1039/c2np20029e.

Ma, W., Zhang, S., Li, Z., Jiang, M., Wang, S., Lu, W., Bi, X., Jiang, H., Zhang, H. and Wei, Z. (2022). Enhancing Protein Function Prediction Performance by Utilizing AlphaFold-Predicted Protein Structures. *Journal of Chemical Information and Modeling*, 62(17), pp.4008–4017. doi:https://doi.org/10.1021/acs.jcim.2c00885.

Majewski, M., Ruiz-Carmona, S. and Barril, X. (2019). An investigation of structural stability in protein-ligand complexes reveals the balance between order and disorder. *Communications chemistry*, 2(1). doi:https://doi.org/10.1038/s42004-019-0205-5.

Malatesta, M., Fornasier, E., Di Salvo, M.L., Tramonti, A., Zangelmi, E., Peracchi, A., Secchi, A., Polverini, E., Giachin, G., Battistutta, R., Contestabile, R. and Percudani, R. (2024). One substrate many enzymes virtual screening uncovers missing genes of carnitine biosynthesis in human and mouse. *Nature Communications*, [online] 15(1). doi:https://doi.org/10.1038/s41467-024-47466-3.

Mani, J., Desy, S., Niemann, M., Chanfon, A., Oeljeklaus, S., Pusnik, M., Schmidt, O., Gerbeth, C., Meisinger, C., Warscheid, B. and Schneider, A. (2015). Mitochondrial protein import receptors in Kinetoplastids reveal convergent evolution over large phylogenetic distances. *Nature Communications*, 6(1). doi:https://doi.org/10.1038/ncomms7646.

McNutt, A.T., Li, Y., Meli, R., Aggarwal, R. and Koes, D.R. (2025). GNINA 1.3: the next increment in molecular docking with deep learning. *Journal of Cheminformatics*, 17(1). doi:https://doi.org/10.1186/s13321-025-00973-x.

Passaro, S., Corso, G., Wohlwend, J., Reveiz, M., Thaler, S., Ram Somnath, V., Getz, N., Portnoi, T., Roy, J., Stark, H., Kwabi-Addo, D., Beaini, D., Jaakkola, T. and Barzilay, R. (2025). Boltz-2: Towards Accurate and Efficient Binding Affinity Prediction. *bioRxiv*. doi:https://doi.org/10.1101/2025.06.14.659707.

Rea, S., Eisenhaber, F., O'Carroll, D., Strahl, B.D., Sun, Z.-W., Schmid, M., Opravil, S., Mechtler, K., Ponting, C.P., Allis, C.D. and Jenuwein, T. (2000). Regulation of chromatin structure by site-specific histone H3 methyltransferases. *Nature*, [online] 406(6796), pp.593–599. doi:https://doi.org/10.1038/35020506.

Rennie, M.L. and Oliver, M.R. (2025). Emerging frontiers in protein structure prediction following the AlphaFold revolution. *Journal of The Royal Society Interface*, 22(225). doi:https://doi.org/10.1098/rsif.2024.0886.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C.L., Ma, J. and Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15), p.e2016239118. doi:https://doi.org/10.1073/pnas.2016239118.

Ruiz-Carmona, S., Schmidtke, P., Luque, F.J., Baker, L., Matassova, N., Davis, B., Roughley, S., Murray, J., Hubbard, R. and Barril, X. (2016). Dynamic undocking and the quasi-bound state as tools for drug discovery. *Nature Chemistry*, 9(3), pp.201–206. doi:https://doi.org/10.1038/nchem.2660.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4), pp.406–425. doi:https://doi.org/10.1093/oxfordjournals.molbev.a040454.

Sankar, S., Preeti, P., Ravikumar, K., Kumar, A., Prasad, Y., Pal, S., Rao, D.N., Savithri, H.S. and Chandra, N. (2023). Structural similarities between SAM and ATP recognition motifs and detection of ATP binding in a SAM binding DNA methyltransferase. *Current Research in Structural Biology*, 6, p.100108. doi:https://doi.org/10.1016/j.crstbi.2023.100108.

Sathyan, K.M., Fachinetti, D. and Foltz, D.R. (2017). α-amino trimethylation of CENP-A by NRMT is required for full recruitment of the centromere. *Nature Communications*, 8(1). doi:https://doi.org/10.1038/ncomms14678.

Sayers, E., Beck, J., Bolton, E., Brister, J., Chan, J., Connor, R., Feldgarden, M., Fine, A., Funk, K., Hoffman, J., Kannan, S., Kelly, C., Klimke, W., Kim, S., Lathrop, S., Marchler-Bauer, A., Murphy, T., O'Sullivan, C., Schmieder, E. and Skripchenko, Y.

(2024). Database resources of the National Center for Biotechnology Information in 2025. *Nucleic Acids Research*, 53(D1). doi:https://doi.org/10.1093/nar/gkae979.

Steverding, D. (2016). Sleeping Sickness and Nagana Disease Caused by *Trypanosoma brucei*. In: *Arthropod Borne Diseases*. Springer, Cham, pp.277–297. doi:https://doi.org/10.1007/978-3-319-13884-8_18.

The UniProt Consortium (2025). *UniProtKB statistics*. [online] uniprot.org. Available at: https://www.uniprot.org/uniprotkb/statistics [Accessed 4 Aug. 2025].

Trott, O. and Olson, A.J. (2009). AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring function, Efficient optimization, and Multithreading. *Journal of Computational Chemistry*, 31(2). doi:https://doi.org/10.1002/jcc.21334.

Unsal, S., Atas, H., Albayrak, M., Turhan, K., Acar, A.C. and Doğan, T. (2022). Learning functional properties of proteins with language models. *Nature Machine Intelligence*, 4(3), pp.227–245. doi:https://doi.org/10.1038/s42256-022-00457-9.

van der Maaten, L. and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, [online] 9(86), pp.2579–2605. Available at: https://www.jmlr.org/papers/v9/vandermaaten08a.html.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I. (2017). *Attention Is All You Need*. [online] arXiv. Available at: https://arxiv.org/abs/1706.03762.

Venna, J. and Kaski, S. (2001). Neighborhood Preservation in Nonlinear Projection Methods: An Experimental Study. In: *Artificial Neural Networks — ICANN 2001*. [online] ICANN 2001. Berlin, Heidelberg: Springer, pp.485–491. Available at: https://doi.org/10.1007/3-540-44668-0_68.

Wagen, C. and Wagen, A. (2025). *The Boltz-2 FAQ*. [online] Rowan Documentation. Available at: https://rowansci.com/blog/boltz2-faq [Accessed 7 Aug. 2025].

Westhorpe, F.G. and Straight, A.F. (2013). Functions of the centromere and kinetochore in chromosome segregation. *Current opinion in cell biology*, [online] 25(3), pp.334–340. doi:https://doi.org/10.1016/j.ceb.2013.02.001.

Wohlwend, J., Corso, G., Passaro, S., Reveiz, M., Leidal, K., Swiderski, W., Portnoi, T., Chinn, I., Silterra, J., Jaakkola, T. and Barzilay, R. (2024). Boltz-1: Democratizing Biomolecular Interaction Modeling. *bioRxiv (Cold Spring Harbor Laboratory)*. doi:https://doi.org/10.1101/2024.11.19.624167.

Wong, J.M. and Eirin-Lopez, J.M. (2021). Evolution of Methyltransferase-Like (METTL) Proteins in Metazoa: A Complex Gene Family Involved in Epitranscriptomic Regulation and Other Epigenetic Processes. *Molecular Biology and Evolution*, 38(12). doi:https://doi.org/10.1093/molbev/msab267.

Yeates, T.O. (2002). Structures of SET Domain Proteins. *Cell*, 111(1), pp.5–7. doi:https://doi.org/10.1016/s0092-8674(02)01010-3.

Yeung, W., Zhou, Z., Mathew, L., Gravel, N., Taujale, R., O'Boyle, B., Salcedo, M., Venkat, A., Lanzilotta, W., Li, S. and Kannan, N. (2023). Tree visualizations of protein sequence embedding space enable improved functional clustering of diverse protein superfamilies. *Briefings in Bioinformatics*, 24(1). doi:https://doi.org/10.1093/bib/bbac619.

Zheng, Y., Koh, H.Y., Yang, M., Li, L., May, L.T., Webb, G.I., Pan, S. and Church, G. (2024). Large Language Models in Drug Discovery and Development: From Disease Mechanisms to Clinical Trials. *arXiv*. doi:https://doi.org/10.48550/arxiv.2409.04481.

Zhonghui, G., Luo, X., Chen, J., Deng, M. and Lai, L. (2023). Hierarchical graph transformer with contrastive learning for protein function prediction. *Bioinformatics*, 39(7). doi:https://doi.org/10.1093/bioinformatics/btad410.

Zhou, N., Jiang, Y., Bergquist, T.R., Lee, A.J., Kacsoh, B.Z., Crocker, A.W., Lewis, K.A., Georghiou, G., Nguyen, H.N., Hamid, M.N., Davis, L., Dogan, T., Atalay, V., Rifaioglu, A.S., Dalkıran, A., Cetin Atalay, R., Zhang, C., Hurto, R.L., Freddolino, P.L.

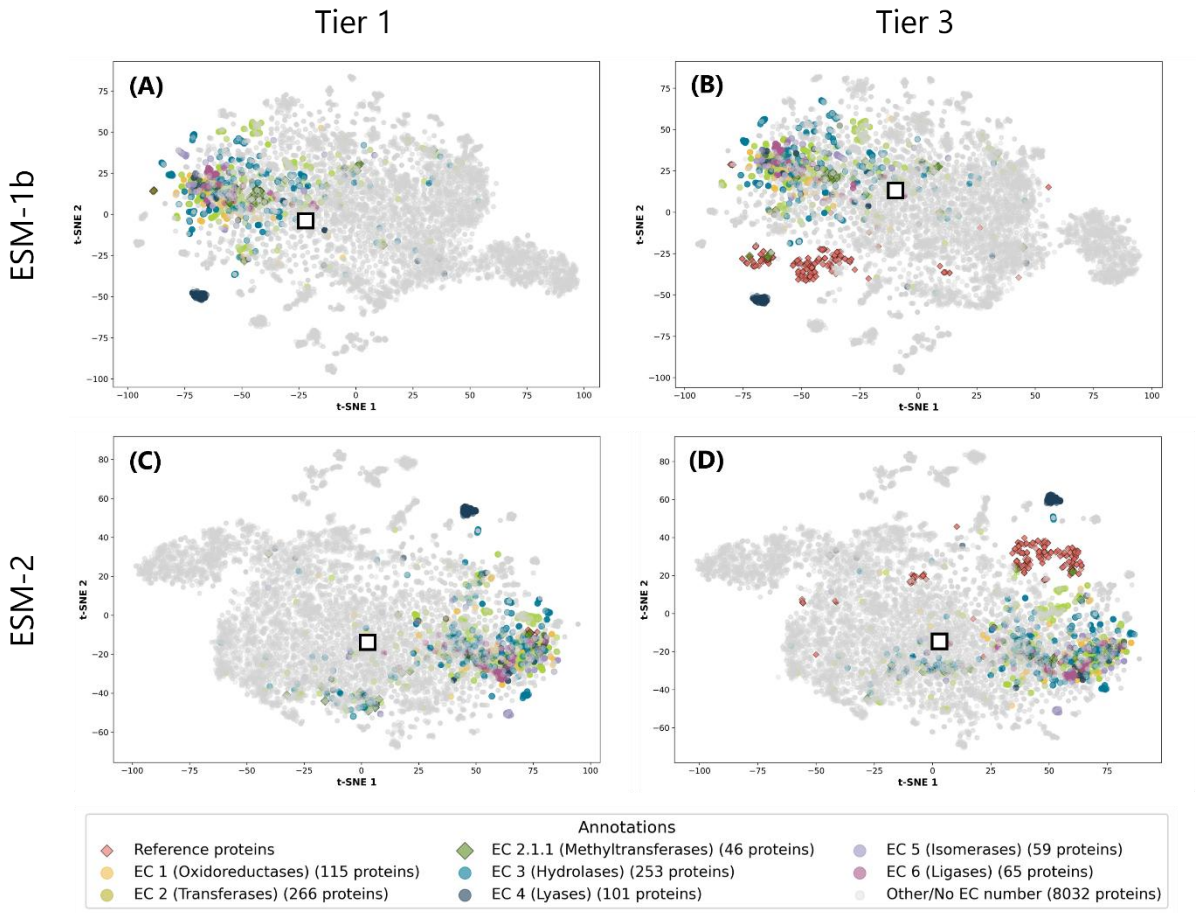
and Zhang, Y. (2019). The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biology*, 20(1). doi:https://doi.org/10.1186/s13059-019-1835-8.

7. Supplementary Materials

Protein ID	Description	Re-docking Vina Affinity (kcal/mol)	Boltz-2 Binding Probability
tr|Q4FKJ2|Q4FKJ2_TRYB2	Methyltransferase	-11.05	0.978
tr|Q4GYA9|Q4GYA9_TRYB2	Arginine N-methyltransferase, putative	-10.95	0.935
tr|Q584S3|Q584S3_TRYB2	Alpha N-terminal protein methyltransferase 1	-10.60	0.962
pdb|6DNZ|C	Chain C, Arginine N-methyltransferase, putative	-10.41	0.949
tr|Q4GZA6|Q4GZA6_TRYB2	Ribosomal RNA-processing protein 8	-10.30	0.964
tr|Q384J5|Q384J5_TRYB2	Uncharacterized protein	-10.14	0.520
tr|Q57U70|Q57U70_TRYB2	Arginine N-methyltransferase, putative	-10.00	0.933
tr|Q57XC2|Q57XC2_TRYB2	Protein arginine methyltransferase NDUFAF7	-10.00	0.808
tr|Q38AF8|Q38AF8_TRYB2	SET domain-containing protein	-9.77	0.933
tr|Q57XE0|Q57XE0_TRYB2	SET domain-containing protein	-9.67	0.808
tr|Q38AF4|Q38AF4_TRYB2	SET domain-containing protein	-9.50	0.877
tr|Q581V5|Q581V5_TRYB2	SET domain-containing protein	-9.46	0.481
tr|Q385Q6|Q385Q6_TRYB2	SET domain-containing protein	-9.44	0.895
pdb|6DNZ|D	Chain C, Arginine N-methyltransferase, putative	-9.40	0.823
tr|Q580C3|Q580C3_TRYB2	SET domain-containing protein	-9.39	0.617
sp|Q582G4|ANM7_TRYB2	Protein arginine N-methyltransferase 7	-9.34	0.929
tr|Q584B2|Q584B2_TRYB2	carnosine N-methyltransferase	-9.25	0.987
tr|Q38BP3|Q38BP3_TRYB2	Arginine N-methyltransferase, putative	-9.16	0.660
tr|Q388V8|Q388V8_TRYB2	SET domain-containing protein	-9.16	0.679
tr|Q382I1|Q382I1_TRYB2	SET domain-containing protein	-9.11	0.577
tr|Q57ZV3|Q57ZV3_TRYB2	SET domain-containing protein	-9.09	0.957
tr|Q57UB6|Q57UB6_TRYB2	SET domain-containing protein	-8.92	0.856
tr|Q57V07|Q57V07_TRYB2	Uncharacterized protein	-8.81	0.736
tr|Q585Q4|Q585Q4_TRYB2	Rubisco LSMT substrate-binding domain-containing protein	-8.75	0.771
tr|Q4GYL8|Q4GYL8_TRYB2	T. brucei spp.-specific protein	-8.54	0.532
tr|Q4GYU9|Q4GYU9_TRYB2	SET domain-containing protein	-8.53	0.354
tr|Q38A07|Q38A07_TRYB2	SET domain-containing protein	-8.51	0.834
pdb|4M38|B	Chain B, Protein arginine N-methyltransferase 7	-8.40	0.892
tr|Q585V9|Q585V9_TRYB2	SET domain-containing protein	-8.31	0.836
tr|Q584E3|Q584E3_TRYB2	SET domain-containing protein	-8.19	0.699
tr|Q57V49|Q57V49_TRYB2	SET domain-containing protein	-8.17	0.890
tr|Q389D9|Q389D9_TRYB2	T. brucei spp.-specific protein	-8.00	0.324
tr|Q57WV8|Q57WV8_TRYB2	SET domain-containing protein	-7.76	0.824
tr|Q57TW0|Q57TW0_TRYB2	SET domain-containing protein	-7.75	0.932
tr|Q57XC0|Q57XC0_TRYB2	SET domain-containing protein	-7.71	0.929
tr|Q57WW4|Q57WW4_TRYB2	SET domain-containing protein	-7.68	0.331

tr|Q583L2|Q583L2_TRYB2	Transposase of Tn10	-7.67	0.413
tr|Q584A8|Q584A8_TRYB2	SET domain-containing protein	-7.56	0.600
tr|Q582H7|Q582H7_TRYB2	SET domain-containing protein	-7.42	0.914
pdb|4V8M|Bs	Chain BP, PROBABLE 60S RIBOSOMAL PROTEIN L14	-7.17	0.566
tr|Q38DZ4|Q38DZ4_TRYB2	SET domain-containing protein	-7.11	0.811
tr|Q582P6|Q582P6_TRYB2	5-demethoxyubiquinone hydroxylase, mitochondrial	-6.98	0.280
tr|Q581S2|Q581S2_TRYB2	Uncharacterized protein	-6.91	0.315
tr|Q582U8|Q582U8_TRYB2	PHD-type domain-containing protein	-6.81	0.186
tr|Q38C25|Q38C25_TRYB2	CFA20 domain-containing protein	-6.77	0.274
tr|Q38CC8|Q38CC8_TRYB2	Uncharacterized protein	-6.65	0.546
tr|Q38EE8|Q38EE8_TRYB2	Nascent polypeptide associated complex subunit, putative	-6.60	0.405
tr|Q4GYQ6|Q4GYQ6_TRYB2	CHORD domain-containing protein	-6.55	0.407
tr|Q585X0|Q585X0_TRYB2	Uncharacterized protein	-6.49	0.466
tr|Q4GYA6|Q4GYA6_TRYB2	SET domain-containing protein	-6.47	0.919
tr|Q385C5|Q385C5_TRYB2	Uncharacterized protein	-6.27	0.295

Supplementary Table 1: Ranked list of predicted SAM-binders. Protein ID refers to the UniProt ID or NCBI ID. The Boltz-2 probability estimates the probability that the ligand (SAM) binds the candidate protein (Wagen and Wagen, 2025). Vina affinity corresponds to the estimated Gibbs free energy change (ΔG) in kcal/mol when SAM is bound.



Supplementary Figure 1: Highlighting Q38BE3 on t-SNE projections of ESM embeddings (perplexity = 30). Projections differ by the choice of ESM model and reference enzyme dataset: **(A)** plots ESM-1b embeddings and reference N-terminal N-methyltransferases (tier 1); **(B)** plots ESM-1b embeddings and reference histone methyltransferases (tier 3); **(C)** plots ESM-2 embeddings and reference N-terminal N-methyltransferases (tier 1); **(D)** plots ESM-2 embeddings and refence histone methyltransferases (tier 3). Each data point represents a protein. Red diamonds indicate reference enzymes, while other colours distinguish Enzyme Commission (EC) numbers of *T. brucei* proteins. Proteins for which an EC number could not be retrieved are coloured grey. White squares with black borders represent protein Q38BE3.

Link to GitHub Repository: https://github.com/B200694/MSc-Project

8. Acknowledgements:

I wish to thank Eugene for his unparalleled enthusiasm for all things chemistry and informatics, his willingness to help, and patience for my incessant questioning. His verve and cheerful profanity was very refreshing and much appreciated.

I also wish to thank Chris for his encyclopaedic knowledge and ability to conjure up decades-past academic papers from thin air. I am very grateful to have had his guidance over the past few months.

Thanks to my parents for their unconditional support from afar, and lastly, to Juny, for keeping me sane.