A Manuscript Microscope 'Sentence Audit' of:

Evotuning protocols for Transformer-based variant effect prediction on multi-domain proteins

Manuscript Source: https://www.biorxiv.org/content/10.1101/2021.03.05.434175v1

Manuscript Authors: Hideki Yamaguchi & Yutaka Saito

Audit Date: 29/03/21 **Audit Identifier:** 5MC1ZJRE59A9K3D **Code Version:** 3.6

Features of the Sentence Audit:

The Sentence Audit combines two complementary proofreading approaches:

1. Each sentence of your text is parsed and displayed in isolation for focused inspection.
2. Each individual sentence is further parsed into Minimal Sentence Components for a deeper review of the clarity, composition and consistency of the language you used.

The Minimal Sentence Components shown are the smallest coherent elements of each sentence of your text as derived from it's conjunctions, prepositions and selected punctuation symbols (i.e. commas, semicolons, round and square brackets).

The combined approaches ensure easier, faster, more effective proofreading.

Comments and Caveats:

• The sentence parsing is achieved using a prototype natural language processing pipeline written in Python and may include occasional errors in sentence segmentation.

• Depending on the source of the input text, the Sentence Audit may contain occasional html artefacts that are parsed as sentences (E.g. "Download figure. Open in new tab").

• Always consult the original research paper as the true reference source for the text.

Contact Information:

To get a Manuscript Microscope Sentence Audit of any other research paper, simply forward any copy of the text to John.James@OxfordResearchServices.com.

All queries, feedback or suggestions are also very welcome.

Research Paper Sections:

The sections of the research paper input text parsed in this audit.

Section No.	Headings	Sentences
Section: 1	Abstract	12
Section: 2	Introduction	15
N/A	--- --- --- --- --- --- ---	0
N/A	--- --- --- --- --- --- ---	0
N/A	--- --- --- --- --- --- ---	0
N/A	--- --- --- --- --- --- ---	0
N/A	--- --- --- --- --- --- ---	0
N/A	--- --- --- --- --- --- ---	0
N/A	--- --- --- --- --- --- ---	0
N/A	--- --- --- --- --- --- ---	0
N/A	--- --- --- --- --- --- ---	0

The Sentence Audit Of The Research Paper

Title

Evotuning protocols for Transformer-based variant effect prediction on multi-domain proteins

S1 [001] Abstract

S1 [002] Accurate variant effect prediction has broad impacts on protein engineering.

Accurate variant effect prediction has broad impacts ...
... on protein engineering.

S1 [003] Recent machine learning approaches toward this end are based on representation learning, by which feature vectors are learned and generated from unlabeled sequences.

Recent machine learning approaches toward this end are based ...
... on representation learning, ...
... by which feature vectors are learned ...
... and generated ...
... from unlabeled sequences.

S1 [004] However, it is unclear how to effectively learn evolutionary properties of an engineering target protein from homologous sequences, taking into account the protein’s sequence-level structure called domain architecture (DA).

However, ...
... it is unclear how ...
... to effectively learn evolutionary properties ...
... of an engineering target protein ...
... from homologous sequences, ...
... taking ...
... into account the protein’s sequence-level structure called domain architecture ...
... (DA).

S1 [005] Additionally, no optimal protocols are established for incorporating such properties into Transformer, the neural network well-known to perform the best in natural language processing research.

Additionally, ...
... no optimal protocols are established ...
... for incorporating ...
... such properties ...
... into Transformer, ...
... the neural network well-known ...
... to perform the best ...
... in natural language processing research.

S1 [006] This article proposes DA-aware evolutionary fine-tuning, or “evotuning”, protocols for Transformer-based variant effect prediction, considering various combinations of homology search, fine-tuning, and sequence vectorization strategies.

This article proposes DA-aware evolutionary fine-tuning, ...

... or “evotuning”, ...

... protocols ...

... for Transformer-based variant effect prediction, ...

... considering various combinations ...

... of homology search, ...

... fine-tuning, ...

... and sequence vectorization strategies.

S1 [007] We exhaustively evaluated our protocols on diverse proteins with different functions and DAs.

We exhaustively evaluated our protocols ...

... on diverse proteins ...

... with different functions ...

... and DAs.

S1 [008] The results indicated that our protocols achieved significantly better performances than previous DA-unaware ones.

The results indicated ...

... that our protocols achieved significantly better performances ...

... than previous DA-unaware ones.

S1 [009] The visualizations of attention maps suggested that the structural information was incorporated by evotuning without direct supervision, possibly leading to better prediction accuracy.

The visualizations ...

... of attention maps suggested ...

... that the structural information was incorporated ...

... by evotuning ...

... without direct supervision, ...

... possibly leading ...

... to better prediction accuracy.

S1 [010] Short descriptions of the authors Hideki Yamaguchi is a PhD candidate at the Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, The University of Tokyo.

Short descriptions ...

... of the authors Hideki Yamaguchi is a PhD candidate ...

... at the Department ...

... of Computational Biology ...

... and Medical Sciences, ...

... Graduate School ...

... of Frontier Sciences, ...

... The University ...

... of Tokyo.

S1 [011] Yutaka Saito, PhD, is a senior researcher at Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST), and a visiting associate professor at The University of Tokyo.

Yutaka Saito, ...
... PhD, ...
... is a senior researcher ...
... at Artificial Intelligence Research Center, ...
... National Institute ...
... of Advanced Industrial Science ...
... and Technology ...
... (AIST), ...
... and a visiting associate professor ...
... at The University ...
... of Tokyo.

S1 [012] Availability https://github.com/dlnp2/evotuning_protocols_for_transformers

Availability ...
... https://github.com/dlnp2/evotuning_protocols_for_transformers

S2 [013] Introduction

S2 [014] A number of mutagenized proteins obtained by evolutionary engineering methods [1] have been reported to demonstrate highly improved functional activity in biomedical research: to name a few, fluorescence [2], signal transduction [3], or base editing capability [4].

A number ...
... of mutagenized proteins obtained ...
... by evolutionary engineering methods ...
... [1] ...
... have been reported ...
... to demonstrate highly improved functional activity ...
... in biomedical research: ...
... to name a few, ...
... fluorescence ...
... [2], ...
... signal transduction ...
... [3], ...
... or base editing capability ...
... [4].

S2 [015] It is well-known that many natural proteins have multiple domains, whose functions are determined by the serial arrangement of domains called domain architecture.

It is well-known ...
... that many natural proteins have multiple domains, ...
... whose functions are determined ...
... by the serial arrangement ...
... of domains called domain architecture.

S2 [016] For example, approximately 65% of eukaryotic proteins are considered to be multi-domain [5].

For example, ...
... approximately 65% ...
... of eukaryotic proteins are considered ...
... to be multi-domain ...
... [5].

S2 [017] Furthermore, various industrially essential proteins, such as artificial antibodies [6] or CRISPR/Cas system-related proteins [7], have multiple domains.

Furthermore, ...
... various industrially essential proteins, ...
... such as artificial antibodies ...
... [6] ...
... or CRISPR/Cas system-related proteins ...
... [7], ...
... have multiple domains.

S2 [018] In protein engineering, a specific domain in a multi-domain protein is often mutagenized (e.g., [8-17]).

In protein engineering, ...
... a specific domain ...
... in a multi-domain protein is often mutagenized ...
... (e.g., ...
... [8-17]).

S2 [019] While useful, the mutagenesis experiments are often costly because multiple iterations of library construction and selection are necessary and they usually require target-specific human knowledge, the latter of which restricts the methods’ generalizability to other proteins.

While useful, ...
... the mutagenesis experiments are often costly ...
... because multiple iterations ...
... of library construction ...
... and selection are necessary ...
... and they usually require target-specific human knowledge, ...
... the latter ...
... of which restricts the methods’ generalizability ...
... to other proteins.

S2 [020] In recent years, machine learning techniques are utilized to predict variant effects for tackling the challenges [18-20].

In recent years, ...
... machine learning techniques are utilized ...
... to predict variant effects ...
... for tackling the challenges ...
... [18-20].

S2 [021] Among them, more attentions are paid to an approach called representation learning, in which features (or descriptors) are directly learned from primary sequences alone [21-23], inspired by natural language processing (NLP) research.

End of Sample Audit