

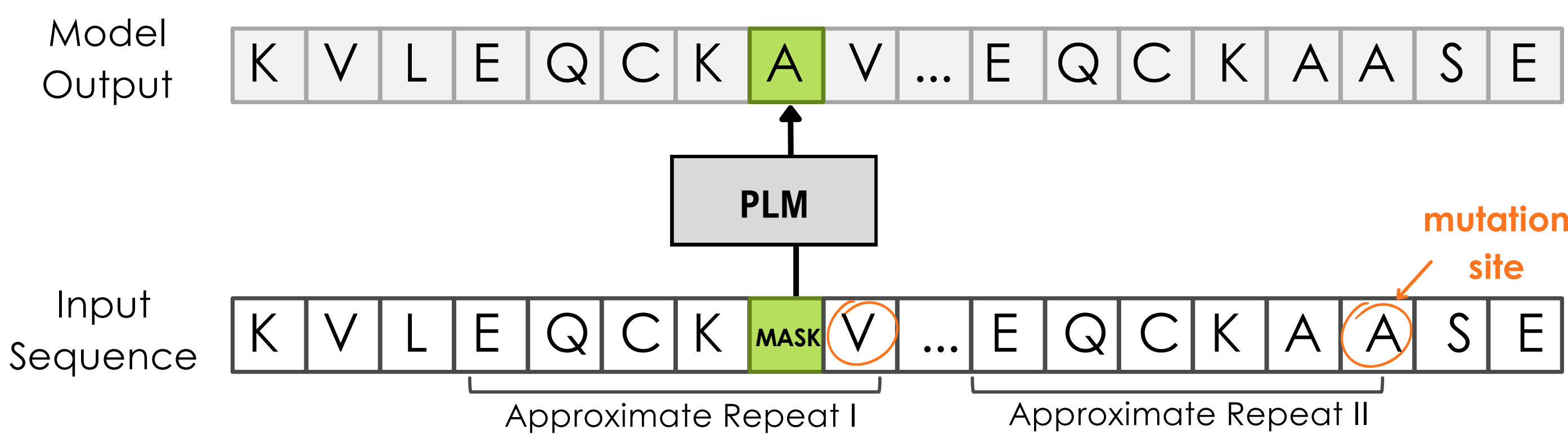
Protein language models detect repeats, even with mutations.

We investigate how.



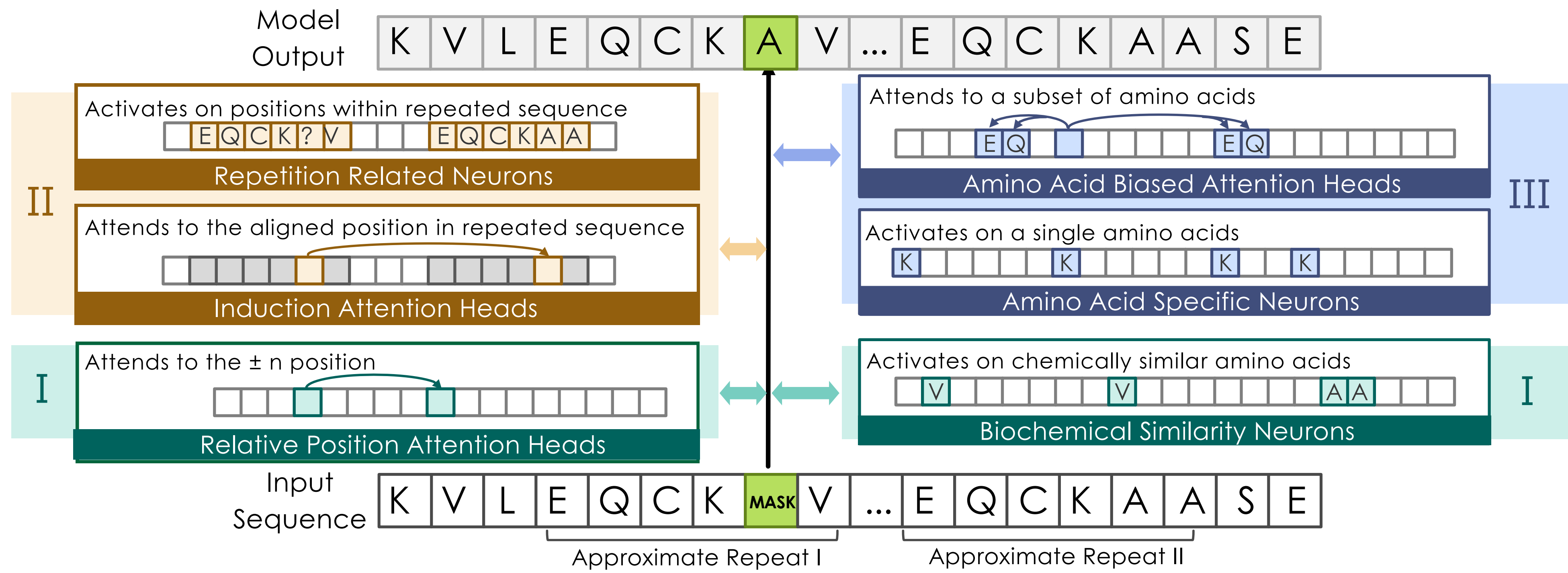
Induction Meets Biology: Mechanisms of Repeat Detection in Protein Language Models
Gal Pomerants, Yaniv Nikankin, Anja Reusch, Tomer Tsaban, Ora Schueler-Furman, Yonatan Belinkov

(1) We formulate **repeat identification as a masked language modeling task.**



We use **three repeat types**: synthetic identical, natural identical, natural approximate.

(2) We find that repeats are identified and completed **by combining an induction-based retrieval mechanism with protein-specific components encoding biochemical similarity.**



(3) We show that many components of the repeat-detection mechanism are **directly interpretable.**

