Protein language models and the semiotic language of Signal Peptides

> Mariana Vitti Rodrigues Claus Emmeche Henrik Nielsen

Are language models learning to read the language of nature?

(Q) What is the role of the computer (language) model in our knowledge about signal peptides?

(H) the semiotic concept of *dicisign*, or *natural proposition* (Stjernfelt 2014), understood as a kind of sign which conveys information, can shed light on the SP phenomenon and why language models may learn to 'read' natural signs.

Summary

- What are Signal Peptides?
- •What is a Dicisign? Can signal peptides be understood as dicisign?
- What are Language Models?
- To which extent can language models read 'the language of nature'?
- •What do language models think about our hypothesis?
- Final Remarks: Open questions

What are Signal Peptides?



Figure 9.7, Cotranslational targeting of secretory proteins to the ER - The Cell - NCBI Bookshelf - Brave

ncbi.nlm.nih.gov/books/NBK9889/figure/A1475/?report=objectonly

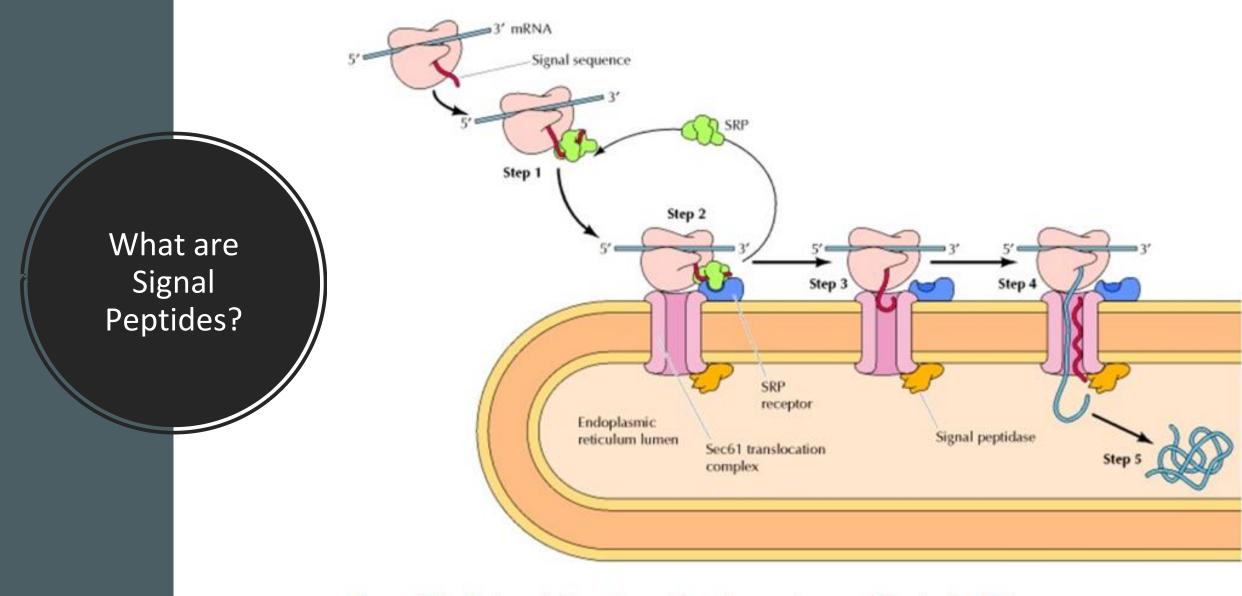
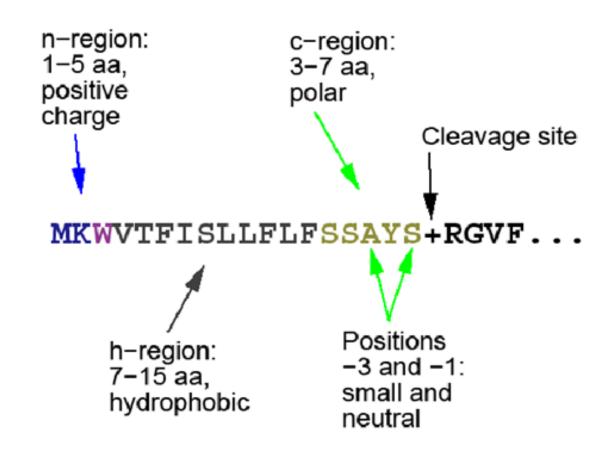


Figure 9.7 Cotranslational targeting of secretory proteins to the ER

What are Signal Peptides?

A **signal peptide** is an N-terminal part of the peptide chain containing a *hydrophobic* region. A typical signal peptide: Human albumin



"[...] proteins have **'intrinsic signals'** that govern their transport and localization in the cell" (Blobel)

In which respect these intrinsic signs can be understood as Dicisigns, i.e., 'signs that say something about something'* (or 'indicates something to be the case'*)?

*Stjernfelt 2011, 2014

What is a Dicisign? Can signal peptides be understood as dicisign?





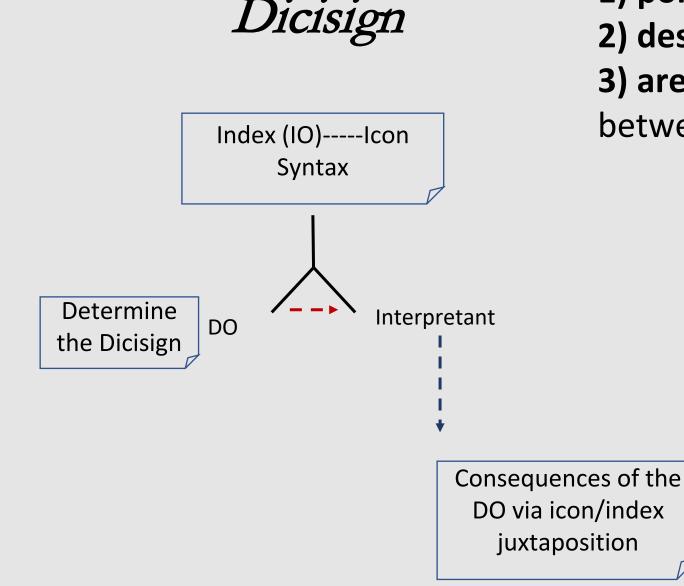
arles Darwin

Dicisign

Dicisigns have a double structure, they:
1) point out an object (its indexical part)
2) describe that object in some way (iconic part).
3) are structured by a syntax (juxtaposition between icon-index)

Dicisigns claim that two signs have the same object; that the two signs form parts or aspects of the Dicisign itself; and that the icon-index syntax holds.

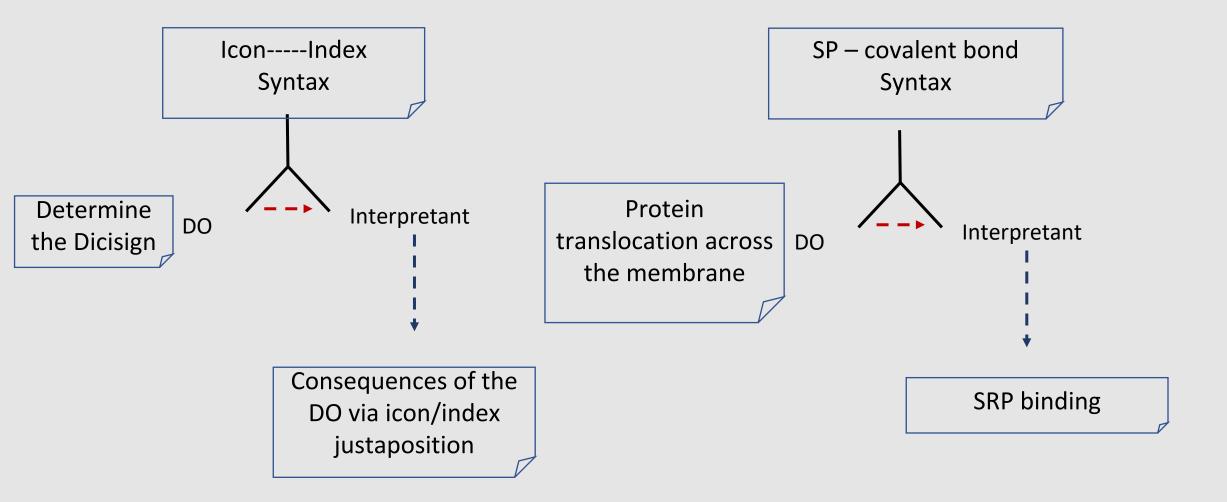
Stjernfelt 2011; Silveira 2008



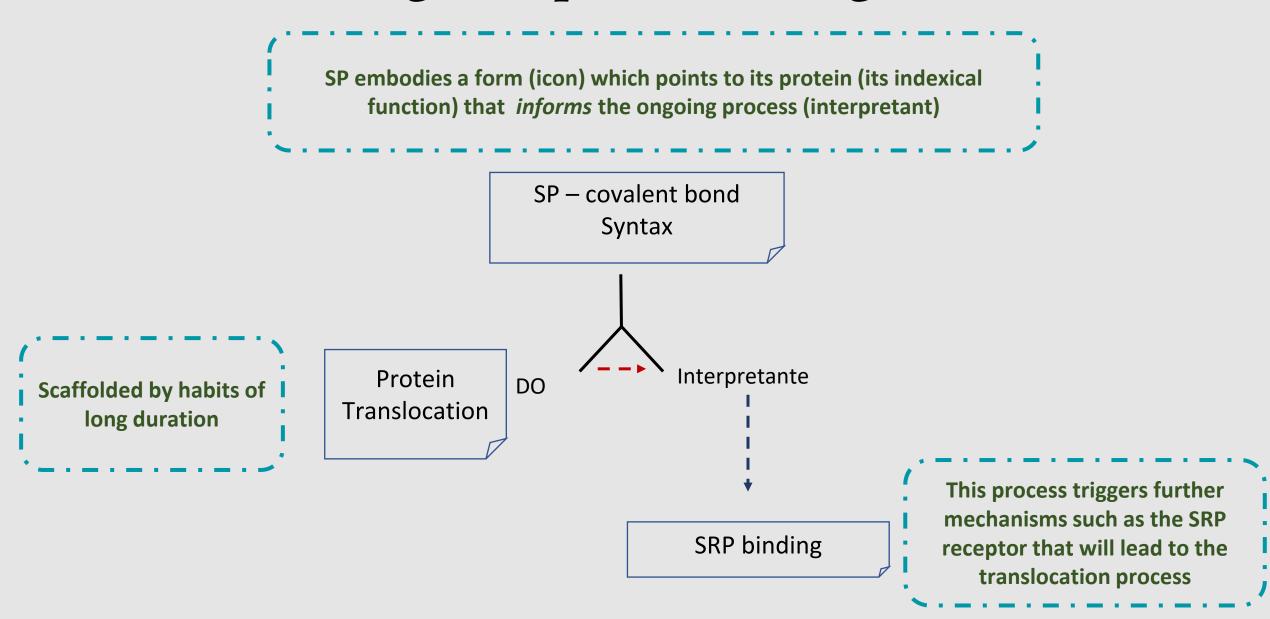
Dicisigns have a double structure, they:
1) point out an object (its indexical part)
2) describe that object in some way (icon).
3) are structured by a syntax (juxtaposition between icon-index)

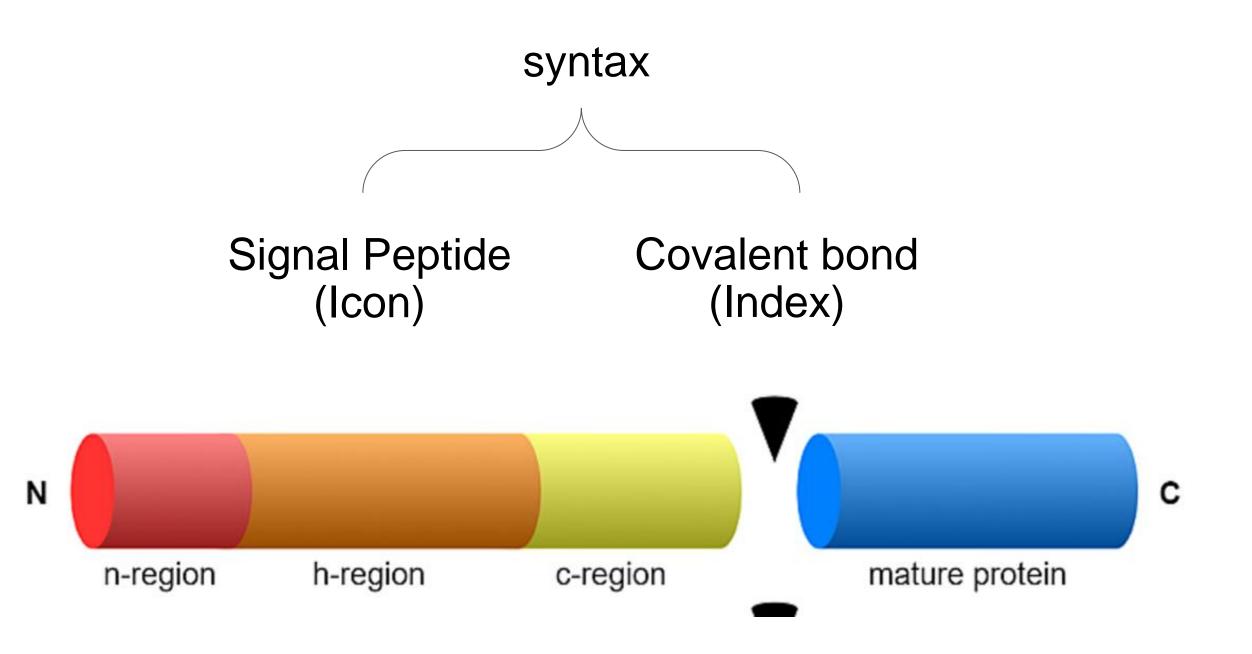
Dicisign

Signal Peptide as Dicisign

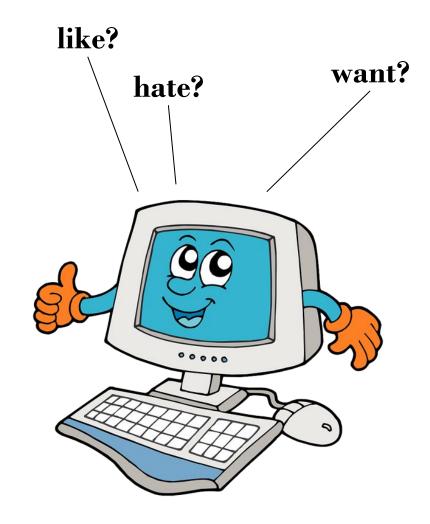


Signal Peptide as Dicisign





What are Language Models (LM)?



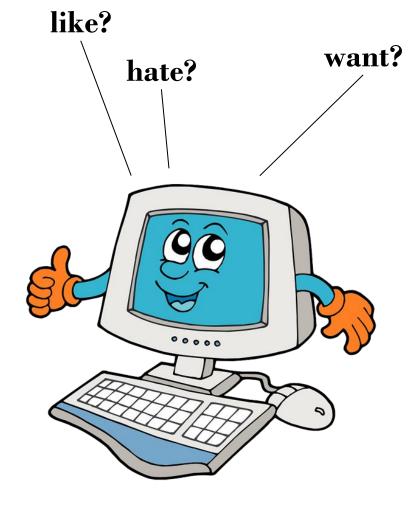
A Language Model (LM)

Mask out random words in human sentences and make the computer predict the missing words:

Do you _____ pizza?

Applications for trained LMs:

- Translation
- Text classification
- Keyword extraction
- Machine-generated answers, e.g. ChatGPT
- Basically, anything that requires the computer to "understand" language

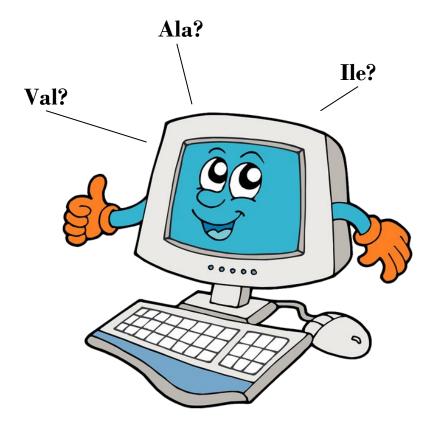


Language models for proteins

Mask out random amino acids in protein sequences and make the computer predict them:

Applications for trained LMs:

- Calculate the probability of a protein
- Generate new proteins
- Create context-dependent representations of amino acids for prediction of
 - structure
 - function
 - location
 - post-translational modifications
 - etc.





BRIEF COMMUNICATION

SignalP 6.0 predicts all five types of signal peptides using protein language models

Felix Teutel¹⁰¹², José Juan Almagro Arm Magnús Halldór Gíslason¹, Silas Irby Pit Søren Brunak¹⁰, Gunnar von Heijne¹⁰

Signal peptides (SPA) are short amino acid see control protein secretion and translocation in all isms. SPs can be predicted from sequence data, algorithms are unable to detect all known types introduce SignalP 6.0, a machine learning mediatenoonis alf five SP types and is applicable to metagenoonis

SPs are short N-terminal ammo acid sequences 0 tetes to the secretory (Sec) pathway in sukaryotes : location across the plasma (inner) membrane in p comprehensive experimental identification of SPs computational prediction of SPs has high relevance cell inology'. SP prediction tools enable identificate that follow the general secretory or twin-arginitie (Tai) pathway and product the position in the sequennal peptidase (SPese) deaves the SPAT, Signal? 5.0 dict Soc substrates cleaved by SPase I (Sec/SPI) or SPII, prokaryotic lipoprotems) and Tat substrates cle I (Tat/SPI)*. However, due to a lack of annotated data unable to detoct Tat substrates cleaved by 3Pase II or processed by SPase III (propilin peptidase, sometime SPase IV'1. Such Sec/SPUE SPs control the translocat pills-like protests, which play a key role in adhesion DNA uptake in prokaryotes'. Furthermore, Signal? regarding the 52⁴ structure, as it cannot define the s N-terminal n-region, the hydropholitic h-region, and i c-region) that underlie the biological function of SPs Here, we present Signal? 6.0, based on protein lan (LMA)** that use information from millions of unant sequences across all domains of life. LMs create sema tations of proteins that capture their biological prope tune. Using these protein representations, Signal? additional types of SPs that previous vensions have b detect while better extrapolating to both proteins dota those used to-create the model and metagemornic data origin. In addition, it is capable of identifying the sub-We comptled a comprehensive dataset of prot that are known to harbor SPs, containing 3,352 3 Sec/SPII, 113 Sec/SPIII, 595 Tat/SPI, 36 Tat/SPII, 1 fular anounces and 2.415 transmetersbrane security Monoover, we defined region-labeling rules accord

Societo to Bioinformatica, Department of Hastith Textmating Science and Engineering, ETH Zurich, Baset, Seitzerland, Nei Science, Livisinship di Capentagen, Capentagen, Denmark, Gammak, Medicine, Eigebaspitate (Capentagen University Augusted Methomatica and Computer Science, Technical Univ Societario University, Biochelut, Sweden, Science for Livie Societario University, Biochelut, Sweden, Science for Livie

NATURE BOTSCHWOLDOV (VIS 40 (2007 2022) 1023-1025) www

DTU Health Tech Department of Health Technology

Research Publications Education Collaboration Services and Products News About

Contact

DTU

Ħ

SHARE ON 🛛 🖌 🖬 🔽

SignalP - 6.0

Prediction of Signal Peptides and their cleavage sites in all domains of life

The Signal F6.0 server predicts the presence of signal peptides and the location of their cleavage sites in proteins from Archaea, Grampositive Bacteria, Gram-negative Bacteria and Eukarya. In Bacteria and Archaea, Signal P6.0 can discriminate between five types of signal peptides:

Chieft for spinknes

Sec/SPI: "standard" secretory signal peptides transported by the Sec translocon and cleaved by Signal Peptidase I (*Lep*) Sec/SPI: lipoprotein signal peptides transported by the Sec translocon and cleaved by Signal Peptidase II (*Lep*) Tat/SPI: Tat signal peptides transported by the Tat translocon and cleaved by Signal Peptidase I (*Lep*) Tat/SPI: Tat lognal peptides transported by the Tat translocon and cleaved by Signal Peptidase I (*Lep*)

That 2 is it is the population signal perposes at unapported by the first fundational man clearle by appoint epidoase it (LSP) SecCSPIII: Plin and plin-like signal perposes at unapported by the Sec transiscont and clearle by Signal Peidoase III (PIID/PibD) Additionally, Signal P G. D predicts the regions as well as of other distinctive features are predicted.

SignaP 6.0 is based on a <u>transformer protein language model</u> with a conditional random field for structured prediction. Behind the Paper: Check out the <u>blog post about the SignaP 6.0 publication</u> in the Nature Portfolio Bioengineering Community. History paper: Check out the <u>blog post about the SignaP Focility</u> in Protein Journal 2019 Eukaryotic proteins: Remember, the presence or absence of a signal peptide is not the whole story about the localization of a protein I with

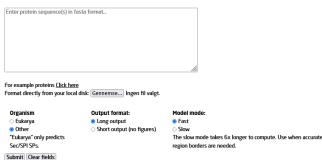
you want to find out more about the sorting of your eukaryotic proteins, try the protein subcellular localization predictor <u>DeepLoc</u>. You may also want to check whether proteins with signal peptides have GPI anchors that keep them attached to the outer face of the plasma membrane using the predictor <u>NetCPI</u>.

Submission Instructions Data Article abstract FAQ Version history Portable Downloads

Submit data

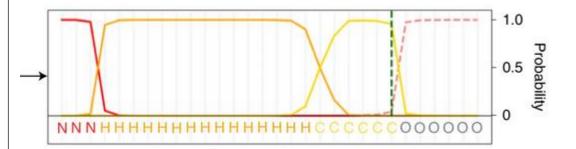
Sequence submission: paste the sequence(s) and/or upload a local file Protein sequences should be not less than 10 amino acids. The maximum number of proteins is 1000. The long output format might timeout for more than 100 entries.

Mirror Use SignalP 6.0 on BioLib if this server is heavily loaded.

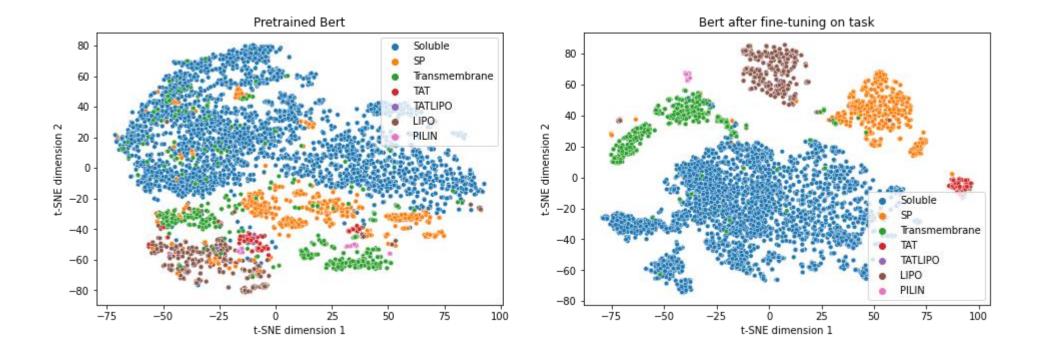


SignalP 6.0

- A program for prediction of signal peptides from amino acid sequences
- Based on a pre-trained protein language model (ProtBert)
- Web server used ~1000 times daily
- Paper published in January 2022
- Cited 486 times (Google Scholar)



Understanding ProtBert



ProtBert has learnt something about the structure of SPs in pretraining — before we told it anything about SPs!

To which extent can language models read 'the language of nature'?

10

0 10 1 0 0 10 1 0 0 10 1 0 0 10 1 0

To which extent can language models read 'the language of nature'?

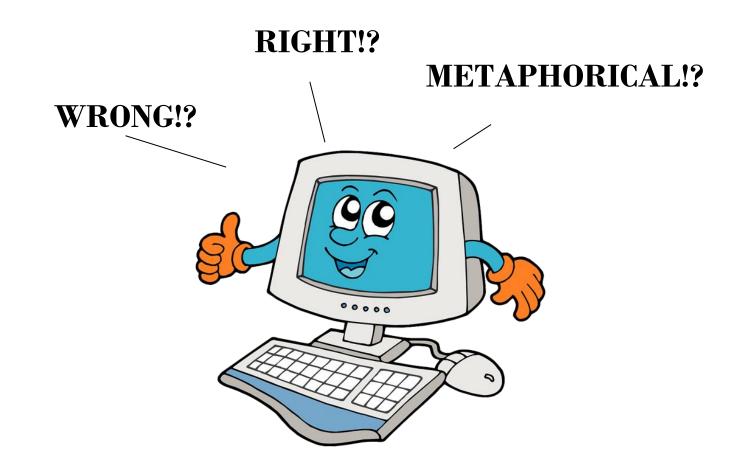
H: the semiotic concept of *dicisign*, or *natural proposition* (Stjernfelt 2014), understood as a kind of sign which conveys information, can shed light on the SP phenomenon and why language models may learn to 'read' natural signs.



- All language models are trained to 'read'/predict some sort of pattern. In this case, the patterns of 'nature' understood as the syntax that structures the subject-predicate relationship of Dicisigns.
- •SignalP 6.0 has shown that language models predict signal peptides by detecting sequences that are likely to function as parts of the syntax of the dicisigns.

"[...] the Syntax depicts the fact involving the object [protein] and quality [SP] corresponding to the S and P parts of the Dicisign"

What do language models think about our hypothesis?



What does the language model **ChatGPT** think about our hypothesis?

Do you think a signal peptide can be viewed as a dicisign?



HE

It could be argued that a signal peptide could be viewed as a dicisign in Peircean semiotics, as it conveys a relationship between a physical or perceptual form (the amino acid sequence of the signal peptide) and a specific meaning (the instruction for the protein to be transported across the membrane). However, it is important to note that the concept of dicisigns is typically applied to linguistic or symbolic signs, whereas signal peptides are physical structures that operate in a biological context. 9

Signal peptides are part of the molecular machinery that regulates protein synthesis and trafficking, and their function is determined by physical and chemical properties of the amino acid sequence, as well as the biological context in which they operate. Therefore, while it may be possible to view signal peptides as a type of dicisign, this would be a somewhat metaphorical or analogical use of the concept, rather than a direct application.

What does the language model built into **Bing** think about our hypothesis?

Do you think a signal peptide can be viewed as a dicisign?

✓ Søger efter "signal peptide dicisign"

/ Genererer svar til dig...

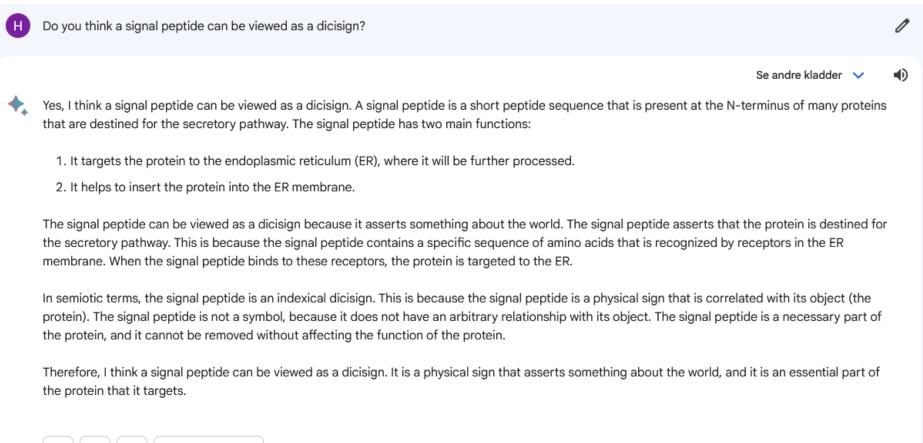
A signal peptide is a short peptide chain that directs the transport of a protein to a specific location in the cell ¹. It is not considered as a dicisign in Peirce's semiotics. Dicisigns are signs that say something about something, and they are not dependent upon conscious stances nor human language ². Signal peptides are not signs that convey information about something else. They are simply sequences of amino acids that direct the protein to its destination ¹.

Få mere at vide: 1. services.healthtech.dtu.dk 2. nature.com

3. sciencedirect.com

4 af 30 🔵

What does the language model **Bard** think about our hypothesis?



🕒 Søg på Google



Open questions

- Is there a sort of 'protein semantics' that is identified by language models, or are there only biochemical processes and statistical patterns?
- Which of the three language models is right about our hypothesis?
- To which extent does a biosemiotic approach to SPs play a relevant role as a source of new knowledge in the growing automation of scientific practice?

Thank you!!!

Mariana Vitti Rodrigues <u>mvittirodrigues@gmail.com</u> São Paulo State University- UNESP

> Claus Emmeche <u>cemmeche@ind.ku.dk</u> University of Copenhagen

Henrik Nielsen <u>henni@dtu.dk</u> Technical University of Denmark

Acknowledgments

São Paulo State University - UNESP FAPESP n. 2023/01408-5 Egenis - University of Exeter

University of Copenhagen

Technical University of Denmark



(not to present, just for our own speculations)

- Is there a sort of 'protein semantics' that is identified by language models, or are there only biochemical processes and statistical patterns?
- Which of the three language models is right about our hypothesis?
- To which extent does a biosemiotic approach to SPs play a relevant role as a source of new knowledge in the growing automation of scientific practice?

- Are the "big language models" really (a) models of language, or (b) models of the patterns of knowledge represented by some language?
- Do they learn to recognize (abductively) patterns, after being trained (inductively) on existing patterns of language, representing knowledge?

Cf. three inferential processes:

- deduction (truth- or knowledge-preserving)
- induction (fallible and knowledge-expanding)
- abduction (fallible and knowledge-expanding)

(not to present, just for our own speculations)

all A are B all B are C *ergo:* all A are C

all A are B

all A are C

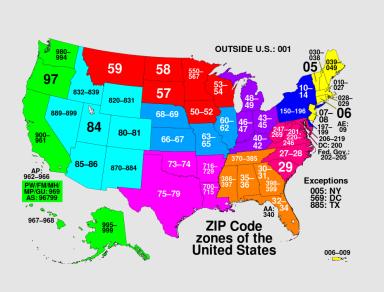
ergo: all B are C

all A are C all B are C *ergo:* all A are B **deductive** pattern, and deductively valid

deductively invalid, but following the pattern of **induction** (at risk making error)

deductively invalid, but following the pattern of **abduction** (at risk making error) — abduction, because: if all B are C, then "all A are B" would explain why all A are C Can a big language model
be seen as a system that is
constructed by training processes involving deduction and induction, lots of errors, and correction of errors (weight adjustments),
then used to process

new inputs (like text prompts, "questions") that are triggering the generation (abductively) of predictive fallible outputs (the "answers")?





a ZIP number as a dicisign within an address code system

point out an object (this letter)
 describe that object (address)
 3) are structured by a syntax (juxtaposition between icon-index)

The postman, knowing the code map, interprets the correct address



Signal Peptide as Dicisign

