



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 Dicsign? Can signal peptides be understood as dicsign?**
- **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?



What are
Signal
Peptides?

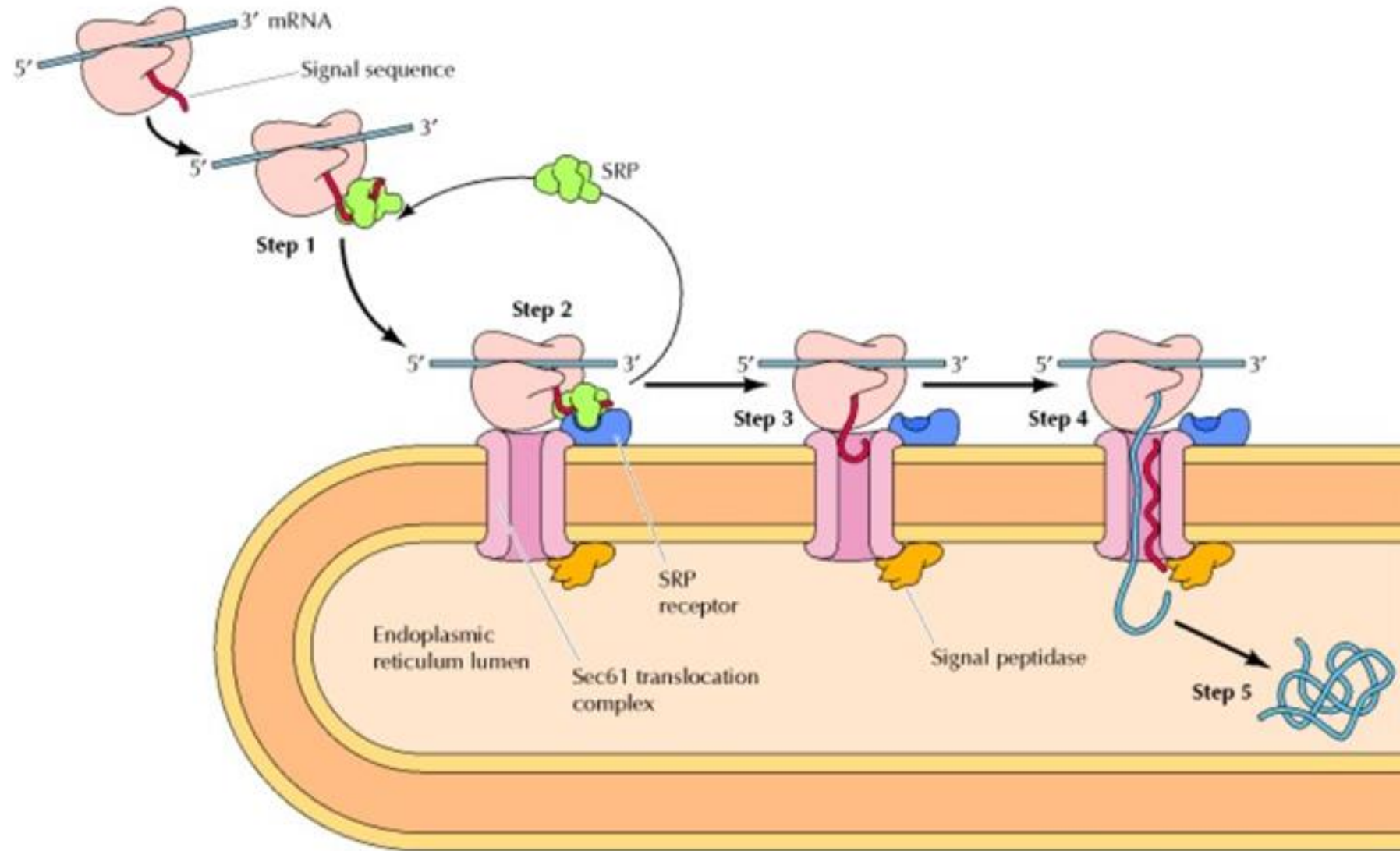
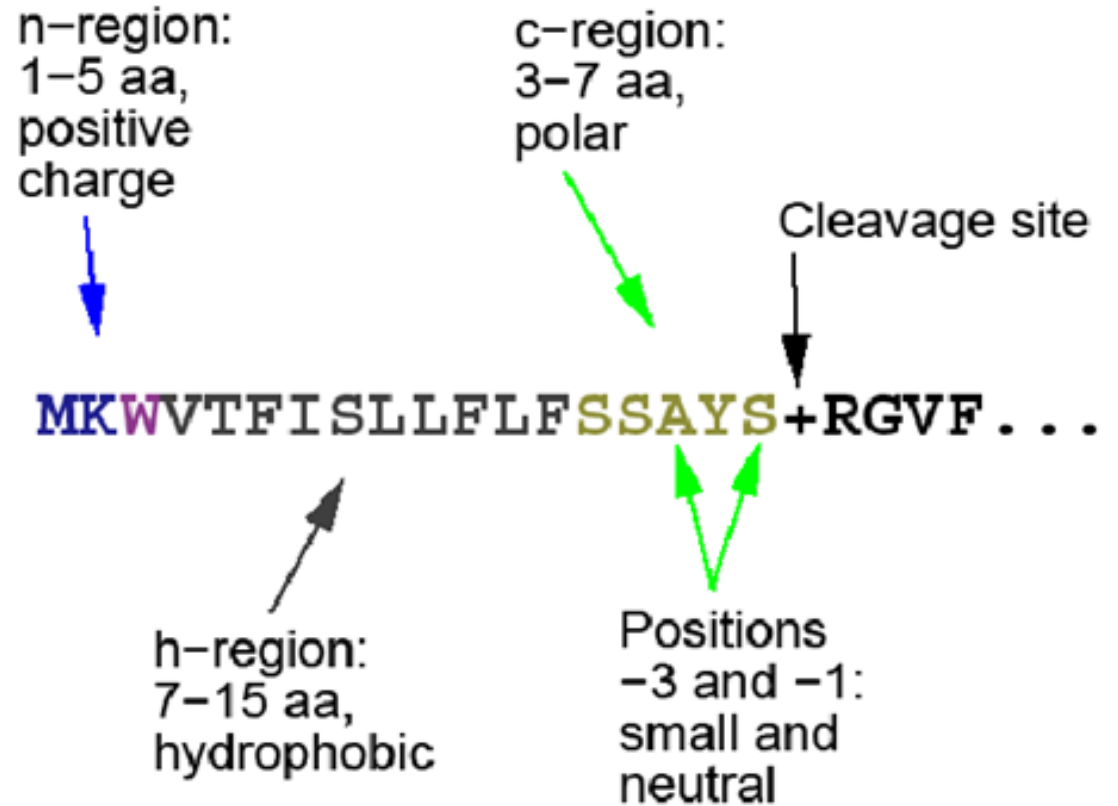


Figure 9.7 Cotranslational targeting of secretory proteins to the ER

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 Dicsign? Can signal peptides be understood as dicsign?





① *Charles Darwin*

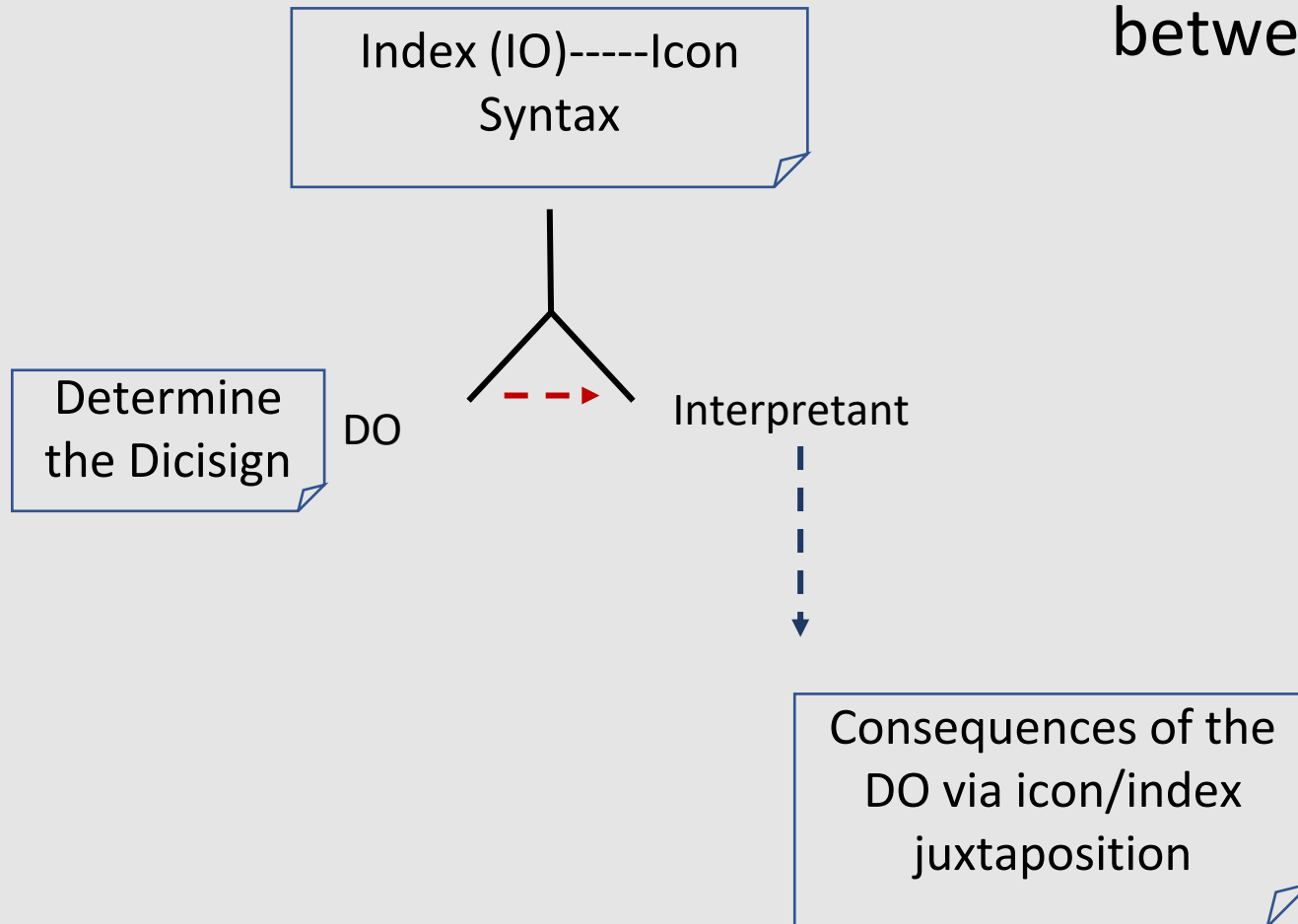
Dicisign

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.

Dicisign



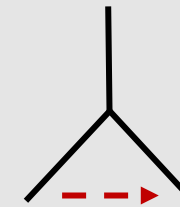
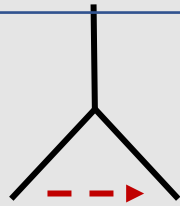
- Dicisigns** have a double structure, they:
- 1) point out an object** (its indexical part)
 - 2) describe that object in some way** (icon).
 - 3) are structured by a syntax** (juxtaposition between icon-index)

Dicisign

Signal Peptide as Dicisign

Icon-----Index
Syntax

SP – covalent bond
Syntax



Determine
the Dicisign

DO

Interpretant

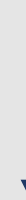
Protein
translocation across
the membrane

DO

Interpretant

Consequences of the
DO via icon/index
justaposition

SRP binding



Signal Peptide as Dicisign

SP embodies a form (icon) which points to its protein (its indexical function) that *informs* the ongoing process (interpretant)

SP – covalent bond
Syntax

Scaffolded by habits of
long duration

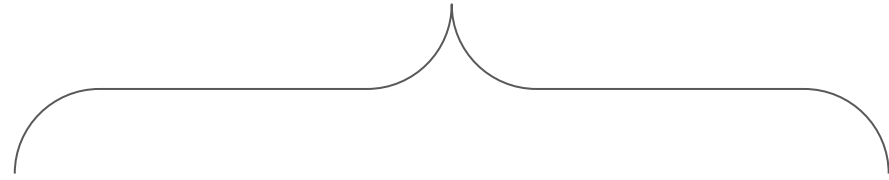
Protein
Translocation DO

Interpretante

SRP binding

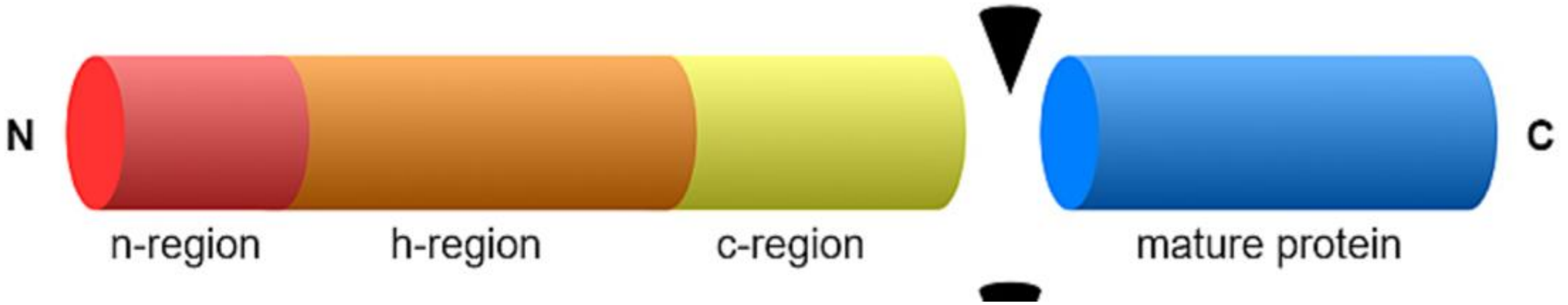
This process triggers further
mechanisms such as the SRP
receptor that will lead to the
translocation process

syntax

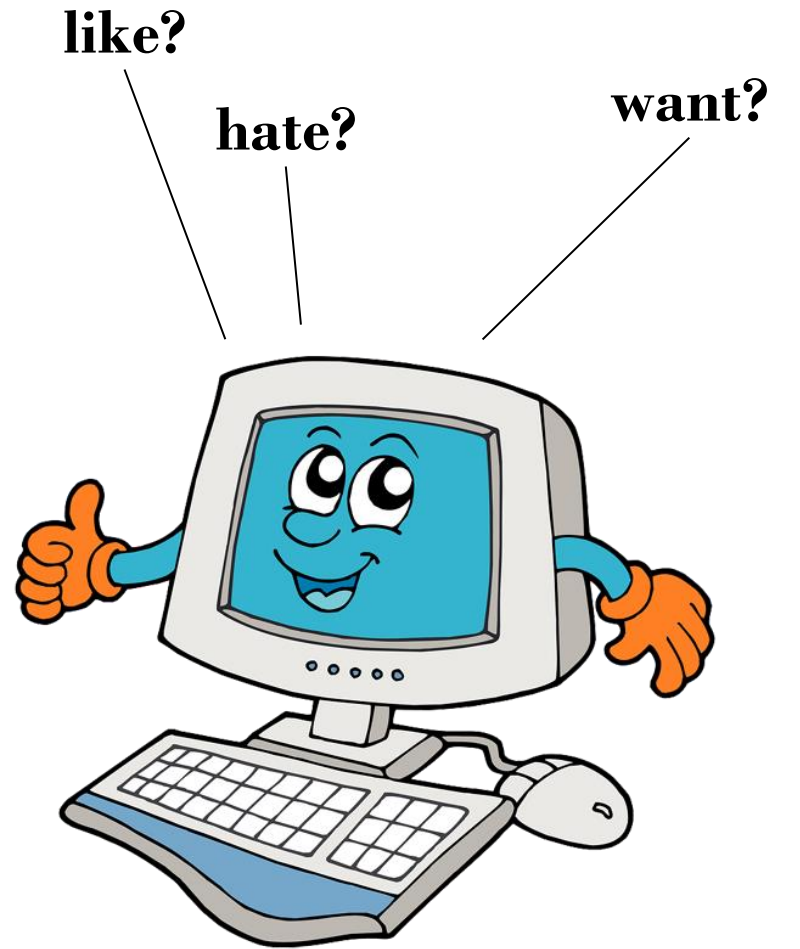


Signal Peptide
(Icon)

Covalent bond
(Index)



What are Language Models (LM)?



A Language Model (LM)

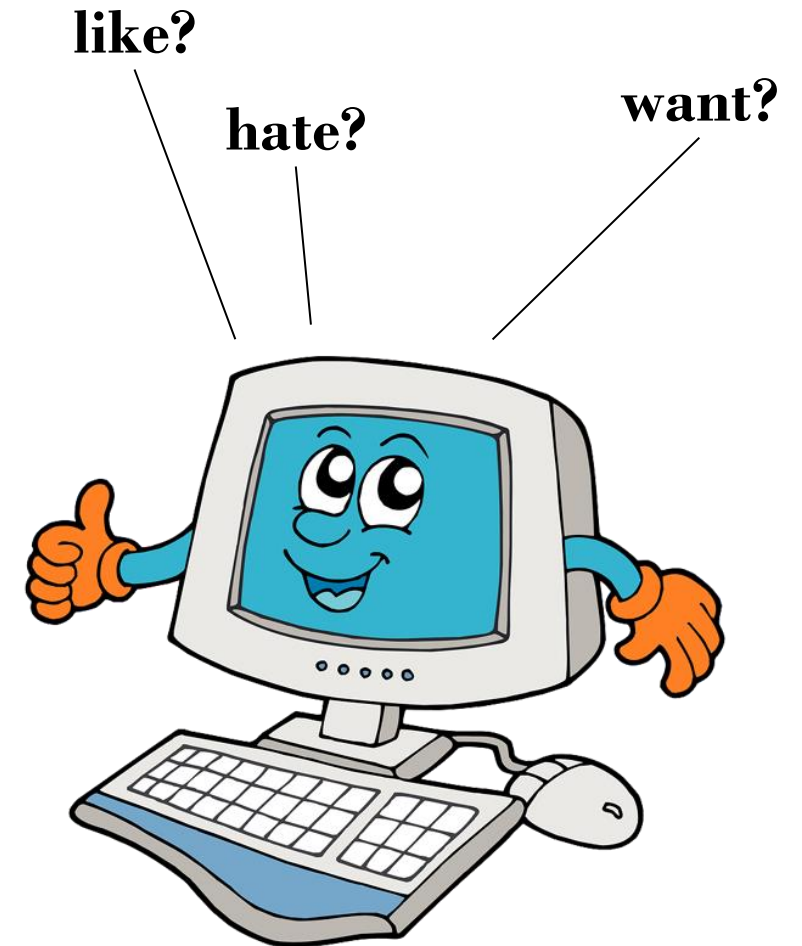
Mask out random words in human sentences and make the computer predict the missing words:



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*



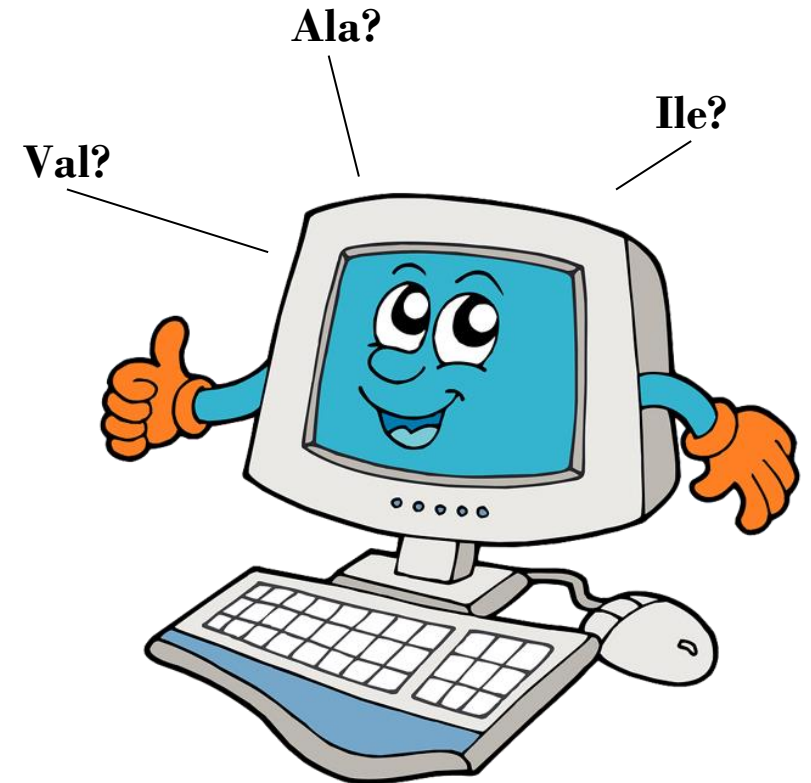
Language models for proteins

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

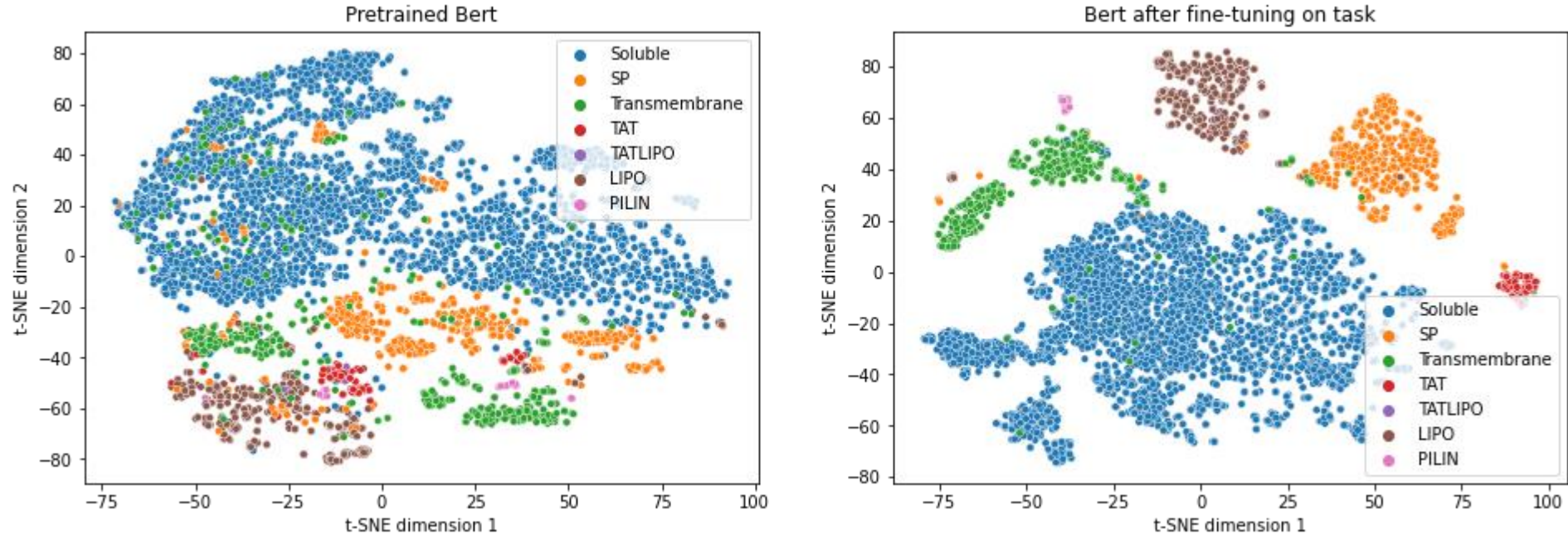
... – Phe – Leu – ____ – Pro – Lys – ...

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.

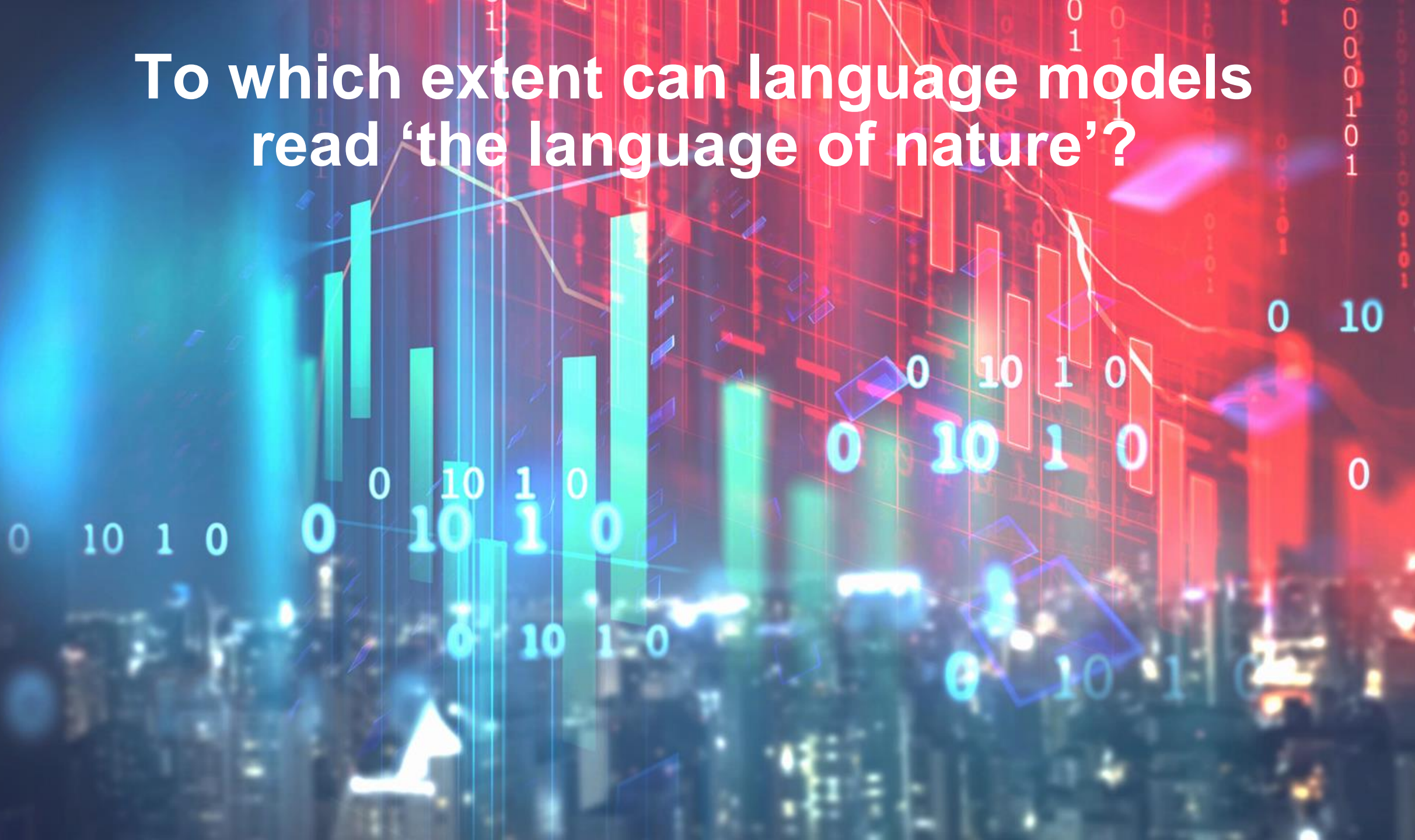


Understanding ProtBert



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'?



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

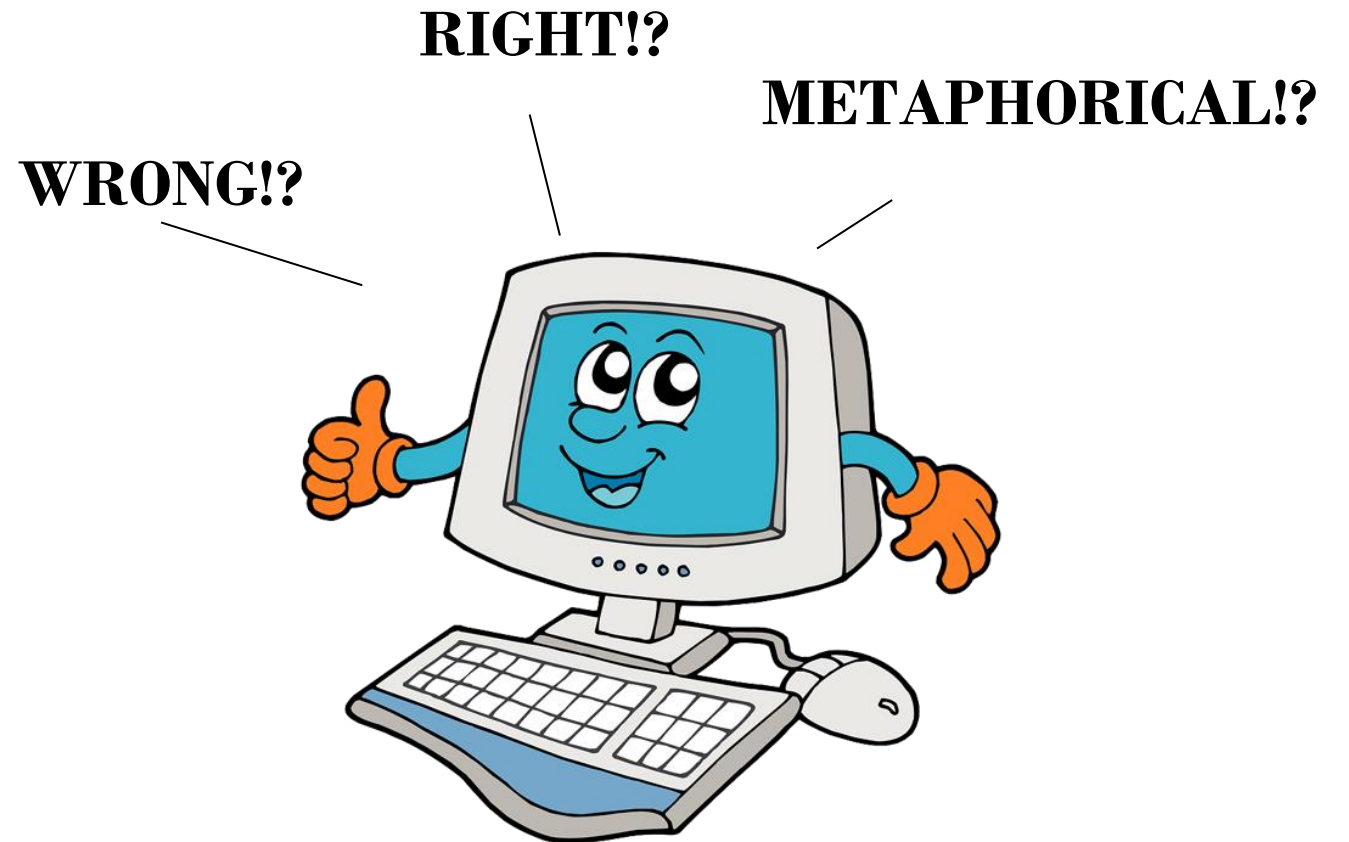
H: the semiotic concept of *dicisign*, or *natural proposition* (Stjernfelt 2014), understood as a kind of sign which conveys information, can shed light on the SP phenomenon and why language models may learn to 'read' natural signs.



- 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?



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.



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 dicensign?


- ✓ Søger efter "signal peptide dicensign"
- ✓ 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 dicensign in Peirce's semiotics. Dicensigns 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](https://services.healthtech.dtu.dk) [2. nature.com](https://nature.com) [3. sciencedirect.com](https://sciencedirect.com)

4 af 30 ●

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

H Do you think a signal peptide can be viewed as a dicensign? 

Se andre kladder  






 Yes, I think a signal peptide can be viewed as a dicensign. A signal peptide is a short peptide sequence that is present at the N-terminus of many proteins that are destined for the secretory pathway. The signal peptide has two main functions:

1. It targets the protein to the endoplasmic reticulum (ER), where it will be further processed.
2. It helps to insert the protein into the ER membrane.

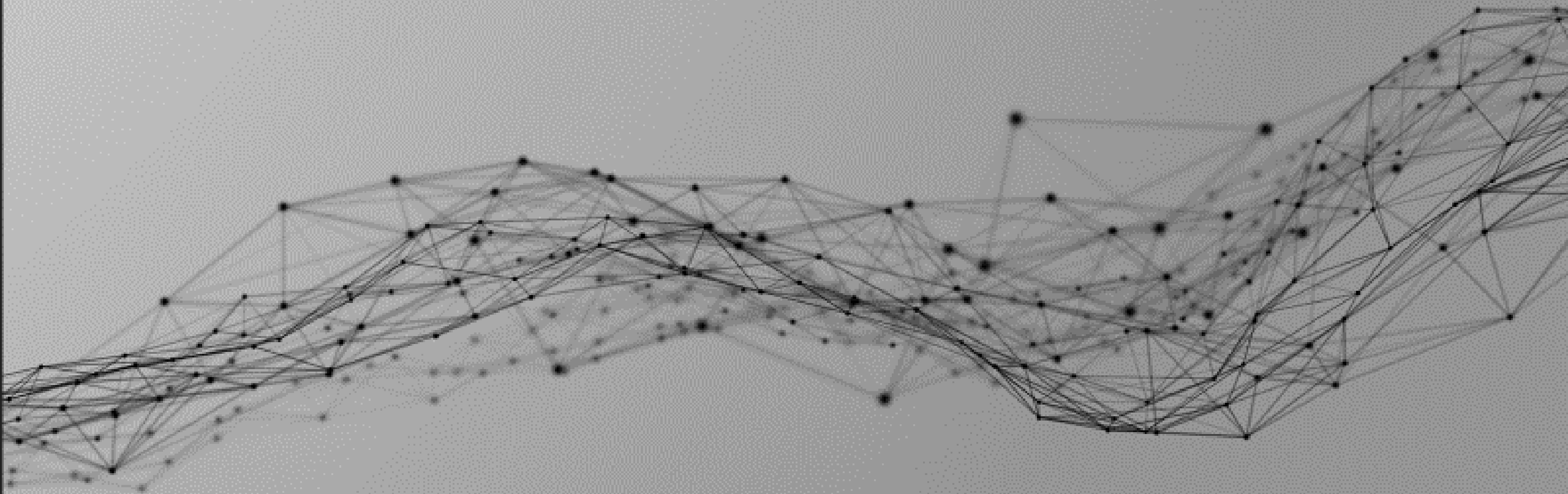
The signal peptide can be viewed as a dicensign because it asserts something about the world. The signal peptide asserts that the protein is destined for the secretory pathway. This is because the signal peptide contains a specific sequence of amino acids that is recognized by receptors in the ER membrane. When the signal peptide binds to these receptors, the protein is targeted to the ER.

In semiotic terms, the signal peptide is an indexical dicensign. This is because the signal peptide is a physical sign that is correlated with its object (the protein). The signal peptide is not a symbol, because it does not have an arbitrary relationship with its object. The signal peptide is a necessary part of the protein, and it cannot be removed without affecting the function of the protein.

Therefore, I think a signal peptide can be viewed as a dicensign. It is a physical sign that asserts something about the world, and it is an essential part of the protein that it targets.

    Søg på Google 

Final Remarks



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

mvittirodrigues@gmail.com

São Paulo State University- UNESP

Claus Emmeche

cemmeche@ind.ku.dk

University of Copenhagen

Henrik Nielsen

henni@dtu.dk

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

} **deductive** pattern, and
deductively valid

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

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

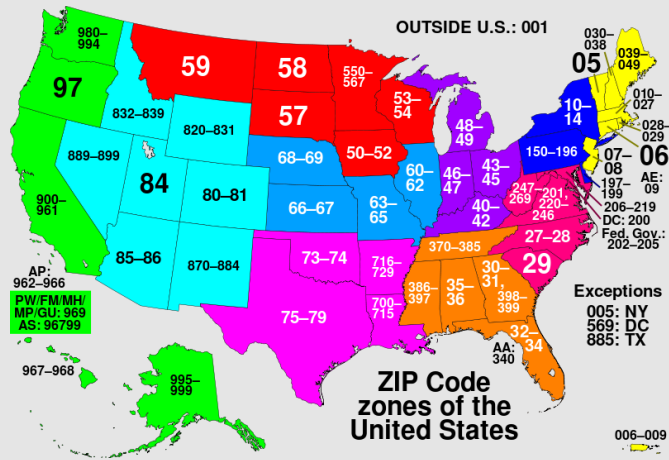
all A are C
all B are C
ergo: all A are B

} 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 decision within an address code system



- 1) point out an object (this letter)
- 2) 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

SP embodies a form (icon) which points to its protein (its indexical function) that *informs* the ongoing process (interpretant)

SP – covalent bond
Syntax

Protein
Translocation

DO

Interpretante

SRP binding

Scaffolded by habits of
long duration

This process triggers further
mechanisms such as the SRP
receptor that will lead to the
translocation process

