BIOMEDICAL

IDENTIFYING IDOMATICITY USING REPRESENTATIONS FROM DEFINITIONS!

● BioLORD

SEMANTICS

MWE IDIOMATICITY

FOR WHO?

CONCEPT EMBEDDING

CLINICAL NLP

BY WHO?

UGENT / IMEC

FRANCOIS REMY

Supervised by K. DEMUYNCK and T. DEMEESTER

FOR WHO?

COLLOCATION SEMANTICS

TRANSLATION TECHNOLOGY

BY WHO?

UMA (MALAGA)

ALFIYA KHABIBULLINA

European Masters in Technology for Translation and Interpreting

PROBLEM STATEMENT

4M?

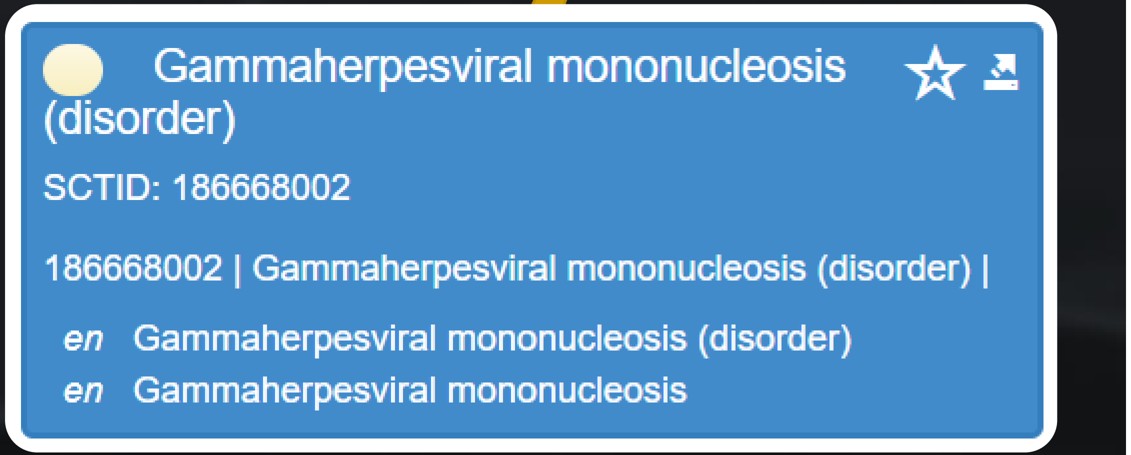
400k

ENGLISH

expert translation

Prioritization?

MULTILING.



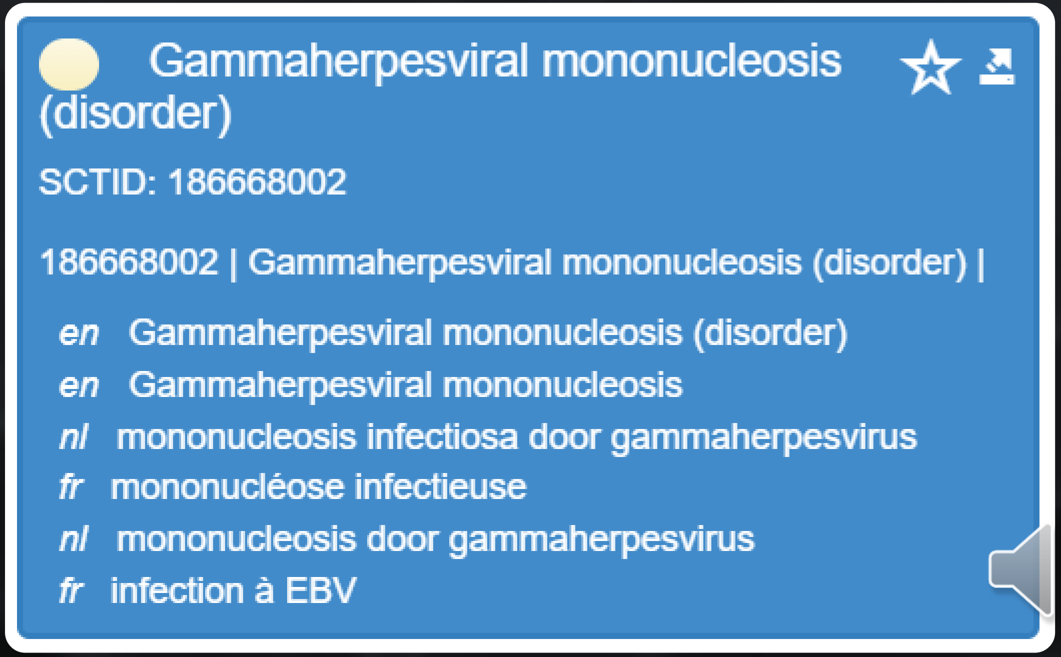
Gammaherpesviral mononucleosis (disorder)

SCTID: 186668002

186668002 | Gammaherpesviral mononucleosis (disorder) |

en Gammaherpesviral mononucleosis (disorder)

en Gammaherpesviral mononucleosis



Gammaherpesviral mononucleosis (disorder)

SCTID: 186668002

186668002 | Gammaherpesviral mononucleosis (disorder) |

en Gammaherpesviral mononucleosis (disorder)

en Gammaherpesviral mononucleosis

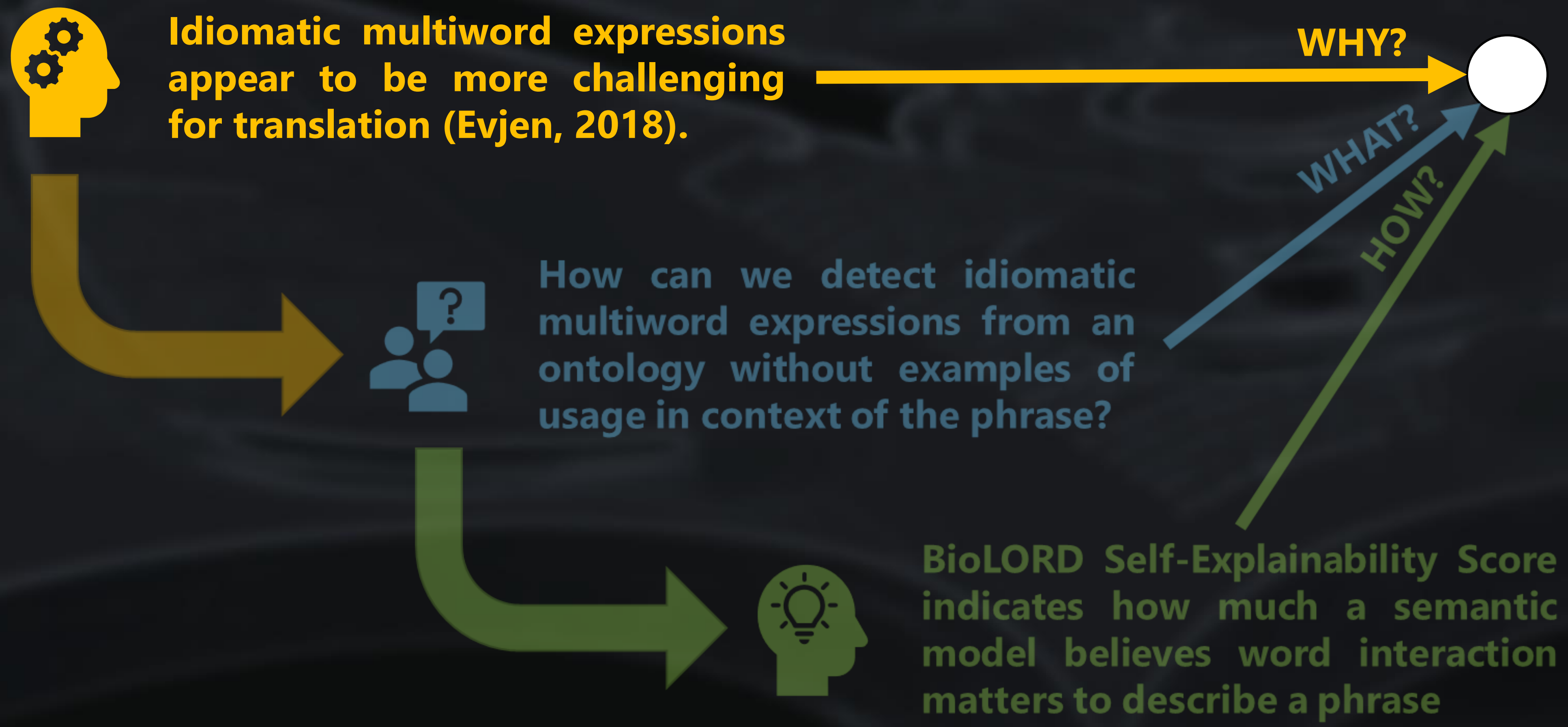
nl mononucleosis infectiosa door gammaherpesvirus

fr mononucléose infectieuse

nl mononucleosis door gammaherpesvirus

fr infection à EBV

IDIOMATIC MWE: WHY?

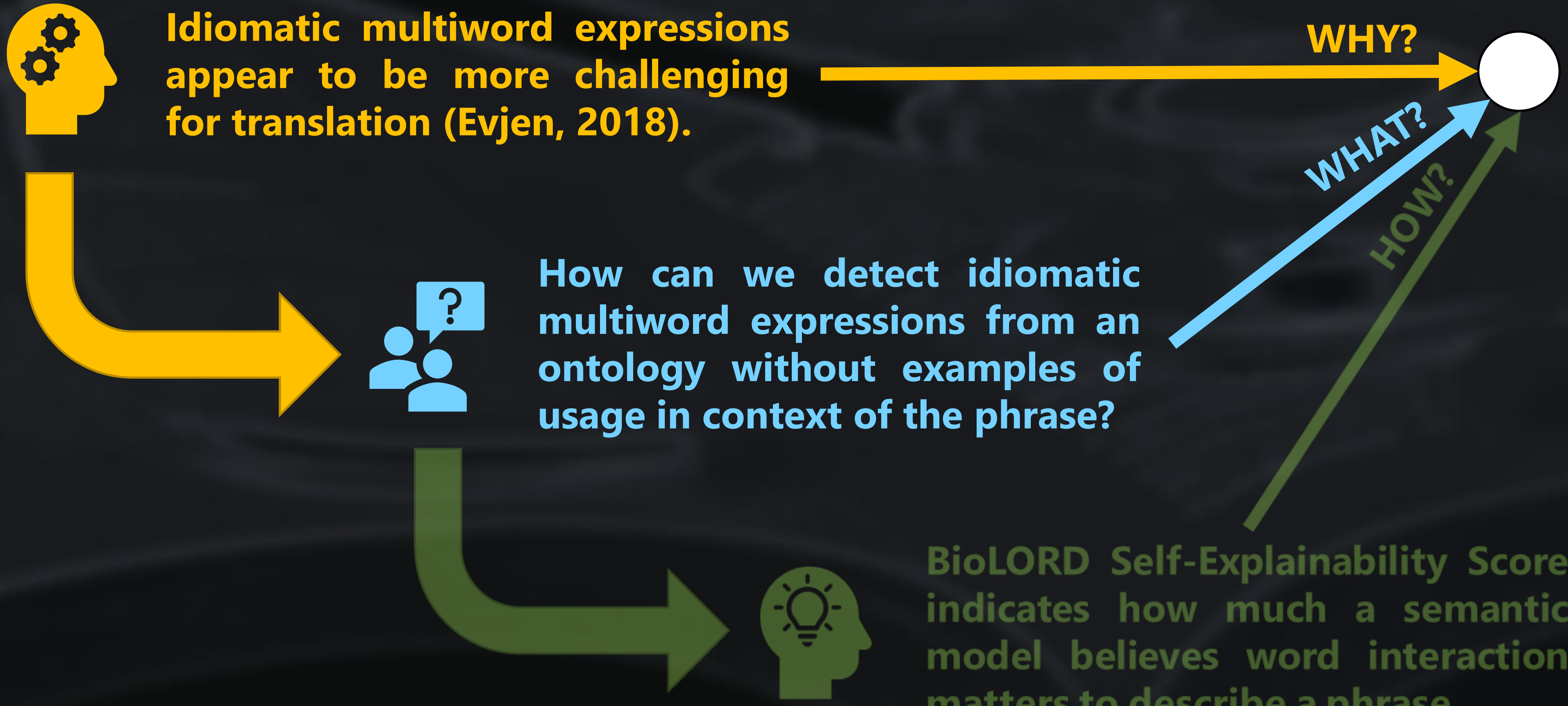


Idiomatic multiword expressions appear to be more challenging for translation (Evjen, 2018).

How can we detect idiomatic multiword expressions from an ontology without examples of usage in context of the phrase?

BioLORD Self-Explainability Score indicates how much a semantic model believes word interaction matters to describe a phrase

IDIOMATIC MWE: WHAT?

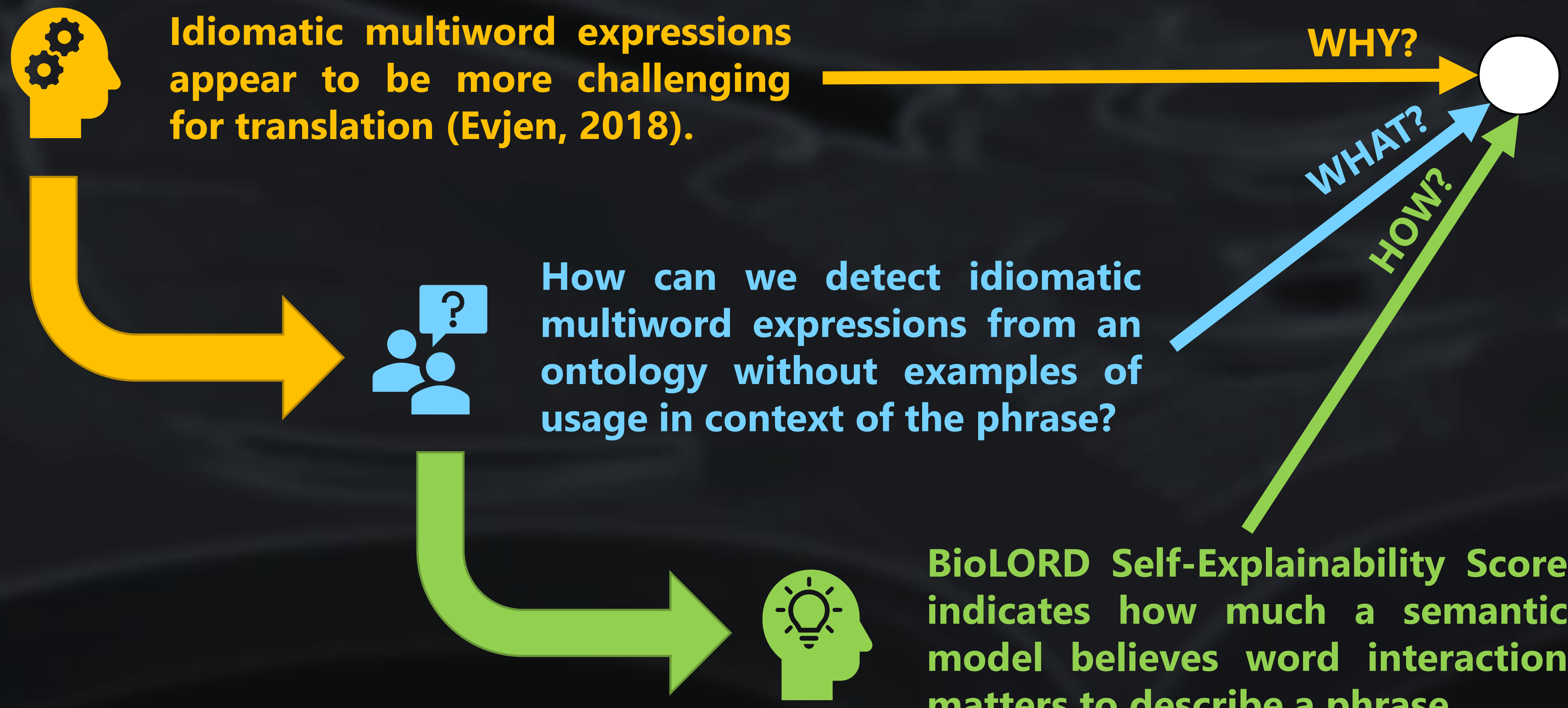


Idiomatic multiword expressions appear to be more challenging for translation (Evjen, 2018).

How can we detect idiomatic multiword expressions from an ontology without examples of usage in context of the phrase?

BioLORD Self-Explainability Score indicates how much a semantic model believes word interaction matters to describe a phrase

IDIOMATIC MWE: HOW?



Idiomatic multiword expressions appear to be more challenging for translation (Evjen, 2018).

How can we detect idiomatic multiword expressions from an ontology without examples of usage in context of the phrase?

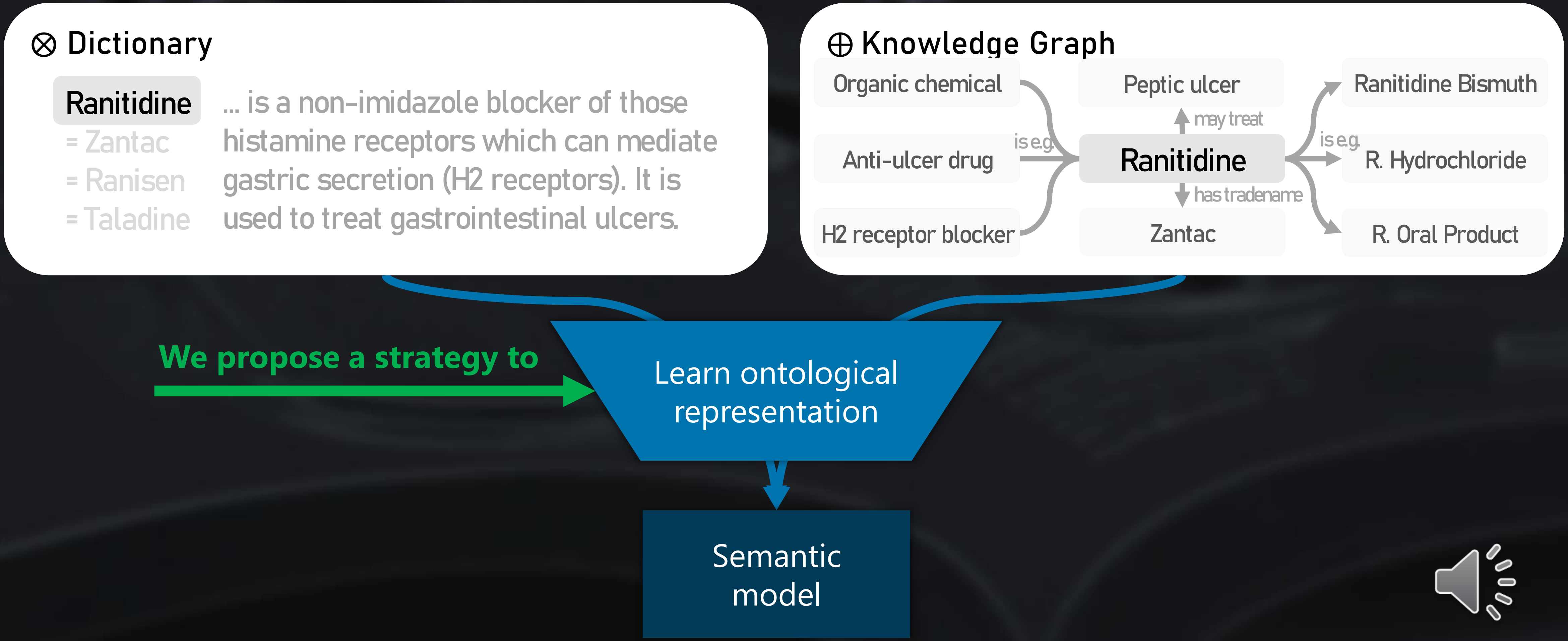
BioLORD Self-Explainability Score indicates how much a semantic model believes word interaction matters to describe a phrase

RELATED WORK: BIOLORD

The representations of concepts and sentences produced by BioLORD are more semantic than state of the art alternatives.

…BUT HOW?

THE BIOLORD TRAINING



⊗ Dictionary

Ranitidine

= Zantac

= Ranisen

= Taladine

... is a non-imidazole blocker of those histamine receptors which can mediate gastric secretion (H2 receptors). It is used to treat gastrointestinal ulcers.

⊕ Knowledge Graph

Organic chemical

Anti-ulcer drug

H2 receptor blocker

is e.g.

Peptic ulcer

may treat

Ranitidine

has tradename

Zantac

is e.g.

Ranitidine Bismuth

R. Hydrochloride

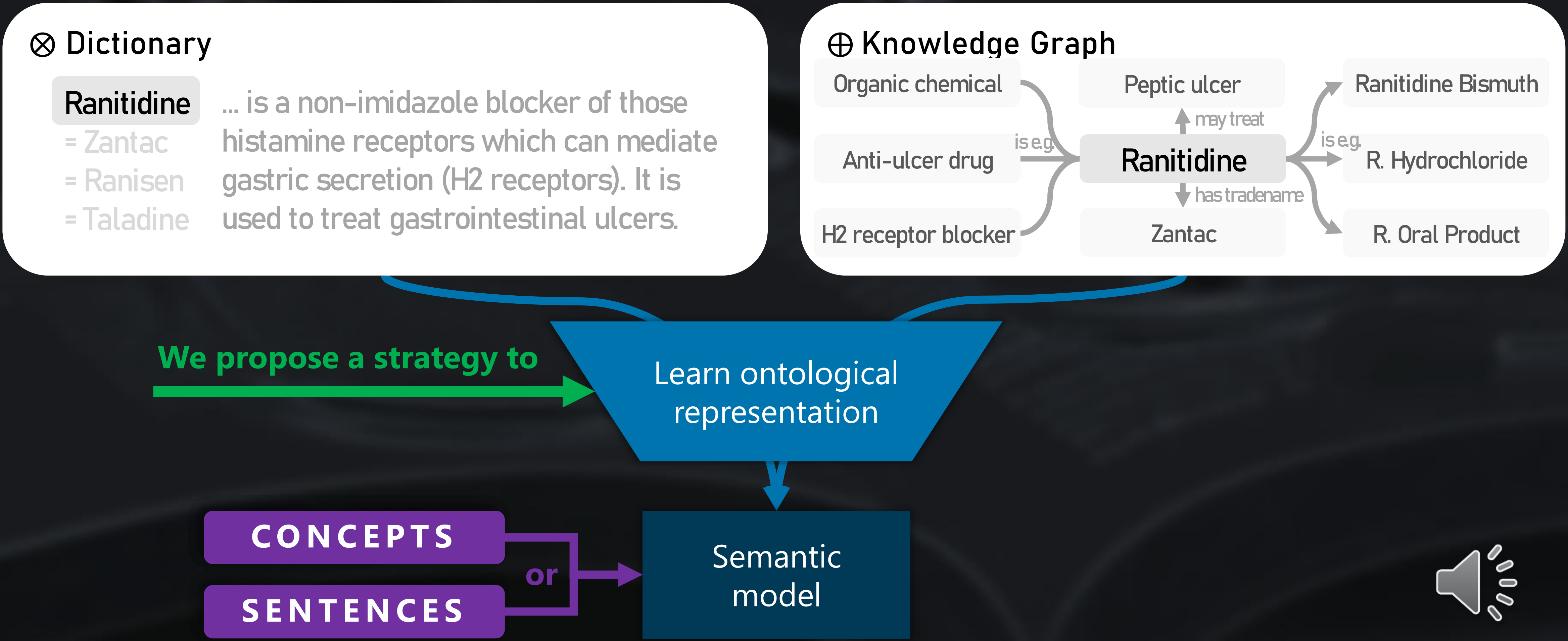
R. Oral Product

We propose a strategy to

Learn ontological representation

Semantic model

THE BIOLORD TRAINING



⊗ Dictionary

Ranitidine
= Zantac
= Ranisen
= Taladine

... is a non-imidazole blocker of those histamine receptors which can mediate gastric secretion (H2 receptors). It is used to treat gastrointestinal ulcers.

⊕ Knowledge Graph

We propose a strategy to

Learn ontological representation

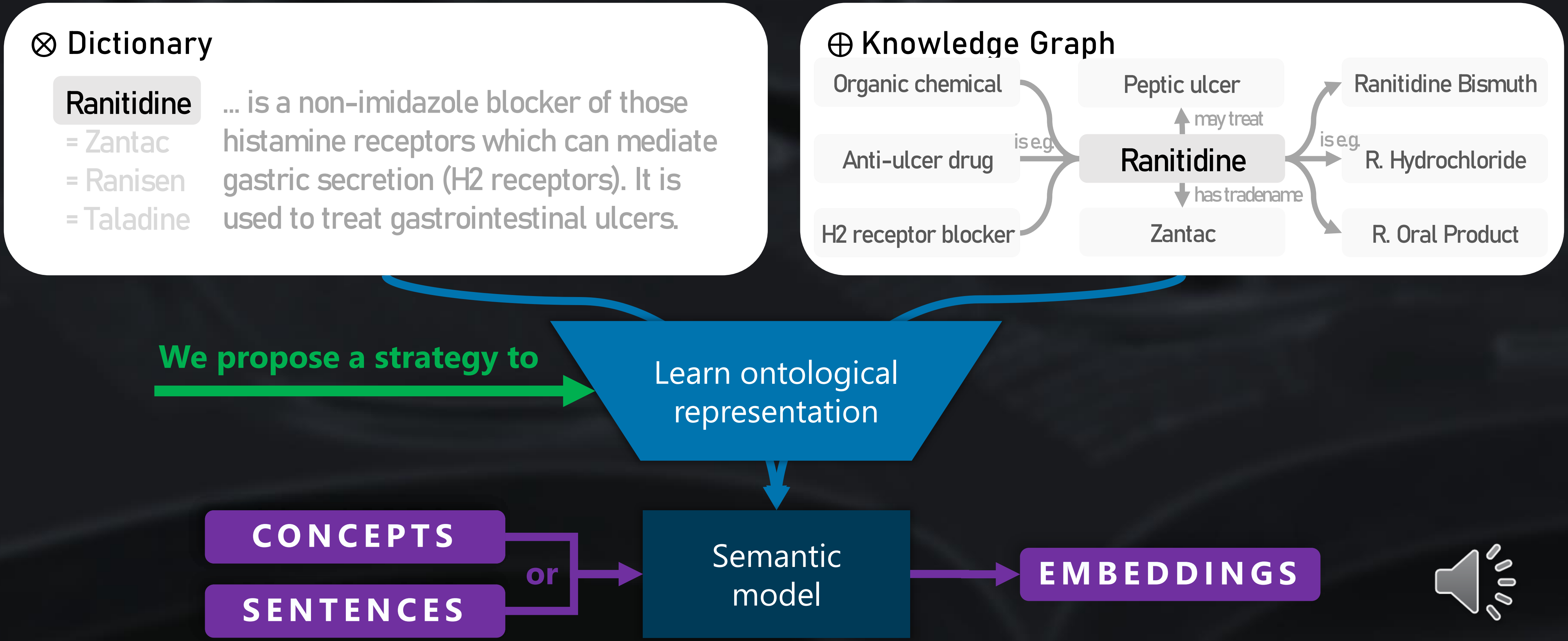
Semantic model

CONCEPTS

or

SENTENCES

THE BIOLORD TRAINING



⊗ Dictionary

Ranitidine

= Zantac

= Ranisen

= Taladine

… is a non-imidazole blocker of those histamine receptors which can mediate gastric secretion (H2 receptors). It is used to treat gastrointestinal ulcers.

⊕ Knowledge Graph

We propose a strategy to

Learn ontological representation

CONCEPTS

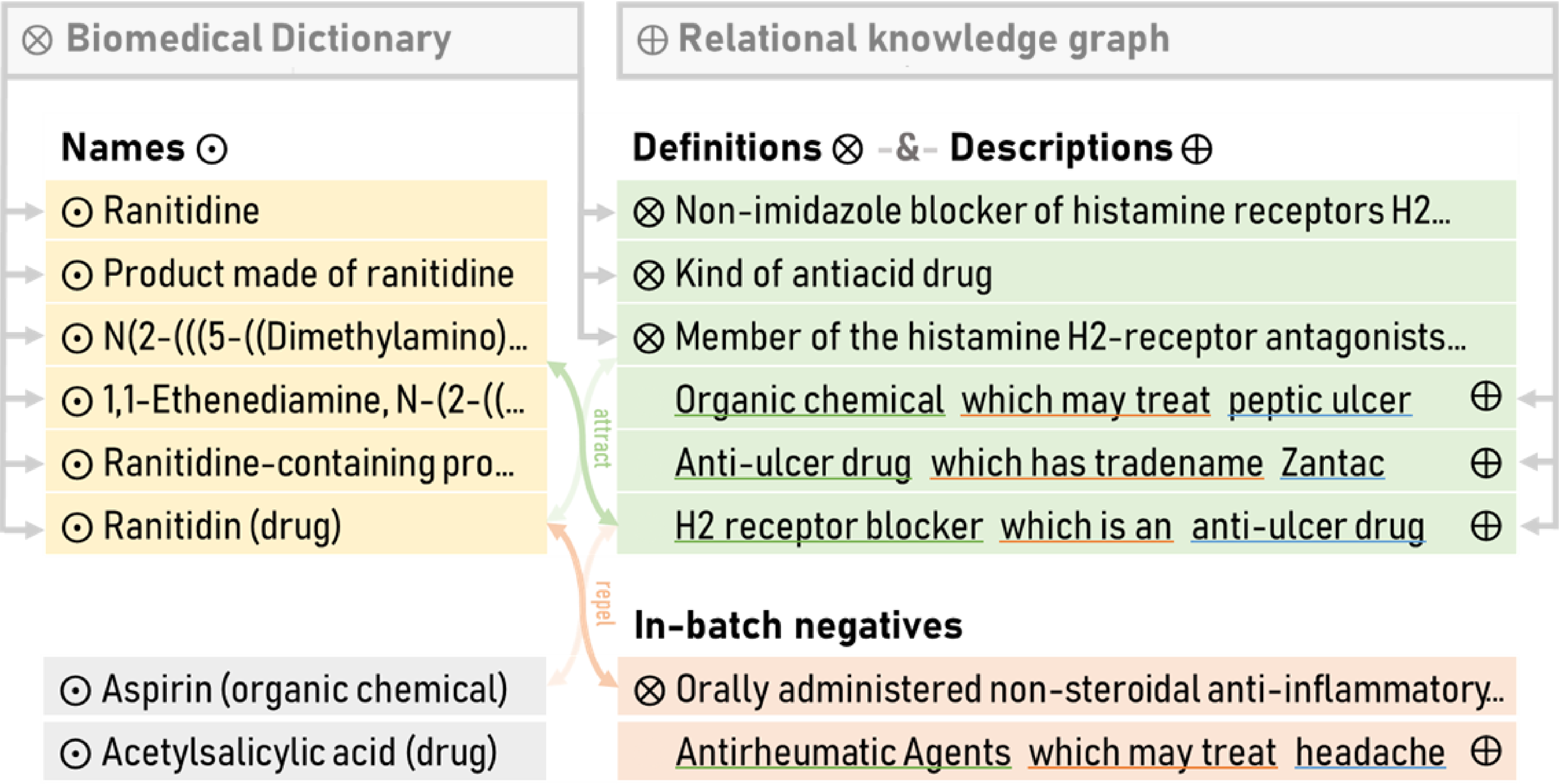
or

SENTENCES

Semantic model

EMBEDDINGS

THE BIOLORD TRAINING



⊗ Biomedical Dictionary

⊕ Relational knowledge graph

Names ⊙

- ⊙ Ranitidine
- ⊙ Product made of ranitidine
- ⊙ N(2-(((5-((Dimethylamino)...
- ⊙ 1,1-Ethenediamine, N-(2-((...
- ⊙ Ranitidine-containing pro...
- ⊙ Ranitidin (drug)

Definitions ⊗ -&- Descriptions ⊕

- ⊗ Non-imidazole blocker of histamine receptors H2...
- ⊗ Kind of antiacid drug
- ⊗ Member of the histamine H2-receptor antagonists...
- Organic chemical which may treat peptic ulcer ⊕
- Anti-ulcer drug which has tradename Zantac ⊕
- H2 receptor blocker which is an anti-ulcer drug ⊕

- ⊙ Aspirin (organic chemical)
- ⊙ Acetylsalicylic acid (drug)

In-batch negatives

- ⊗ Orally administered non-steroidal anti-inflammatory...
- Antirheumatic Agents which may treat headache ⊕

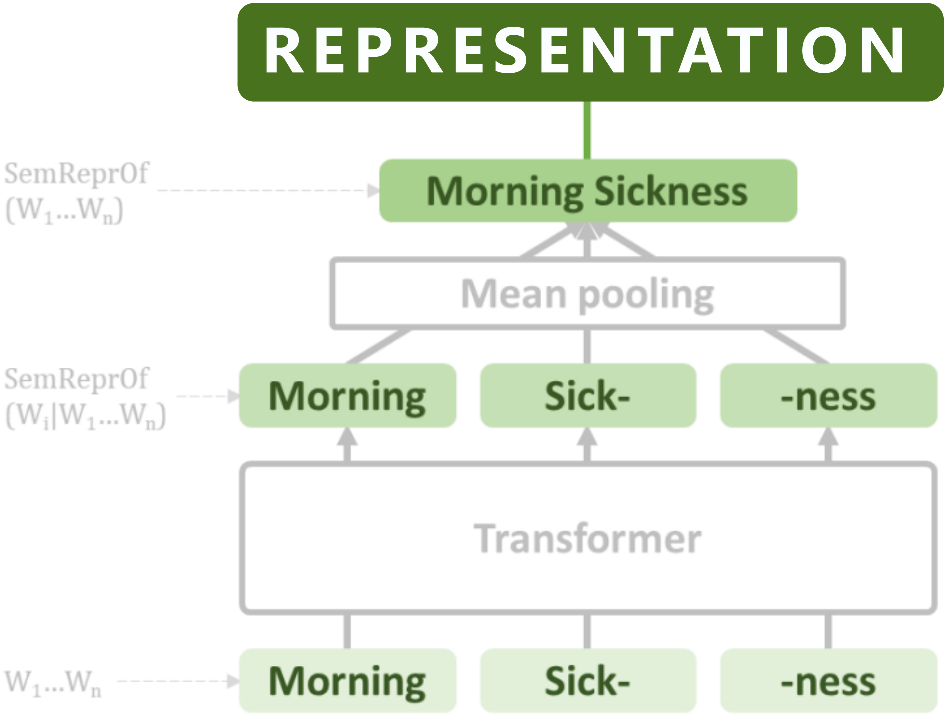
METHODOLOGY

BioLORD Self-Explainability Score indicates how much a semantic model believes word interaction matters to describe a phrase...

...BUT HOW?

SEMANTIC REPRESENTATION

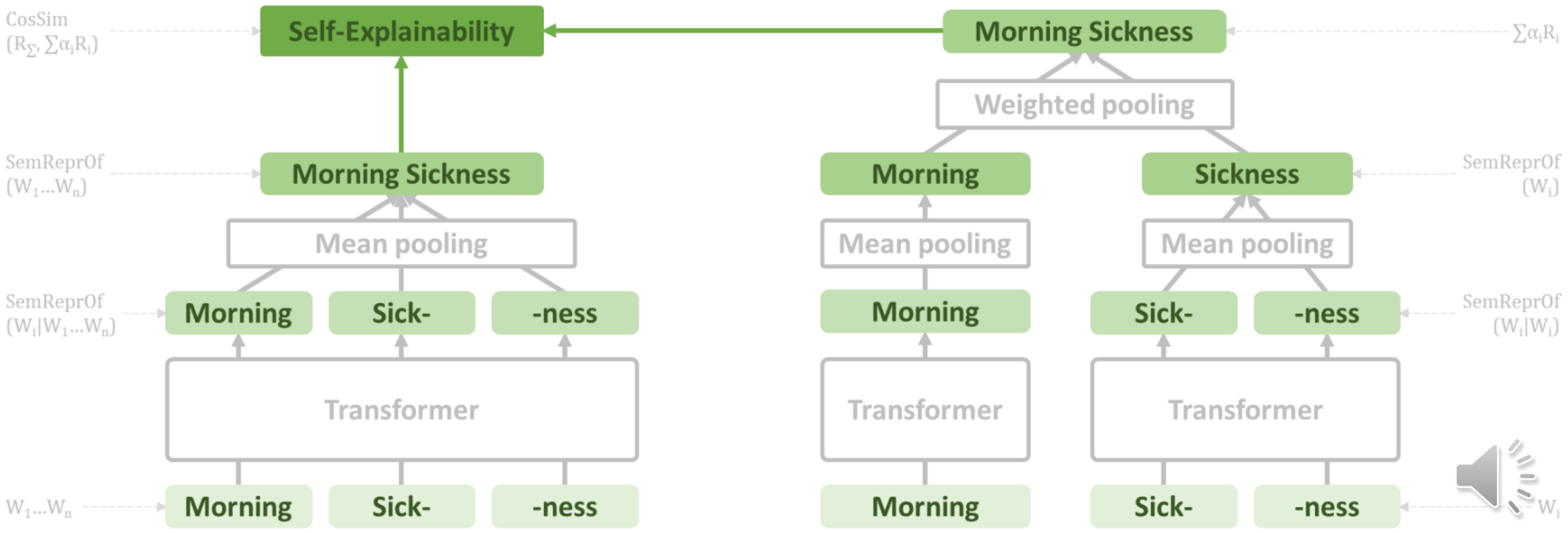
We first compute a semantic representation for the concept



The semantic representation of concepts is computed by averaging the embedding of its tokens, after interactions between tokens have been taken into consideration by the Transformer model

SELF-EXPLAINABILITY SCORE

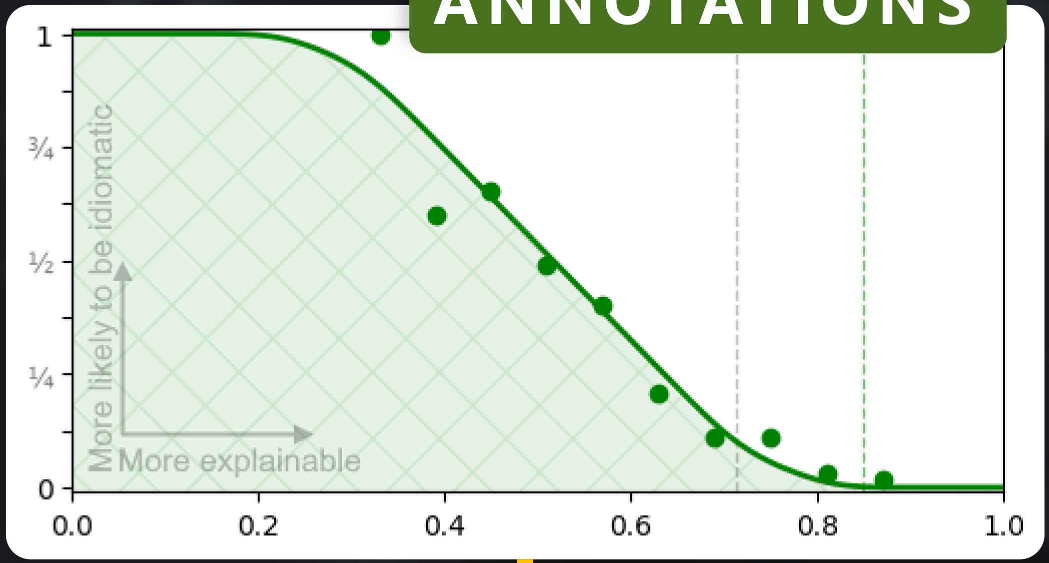
We can compare it to an average that ignores interactions

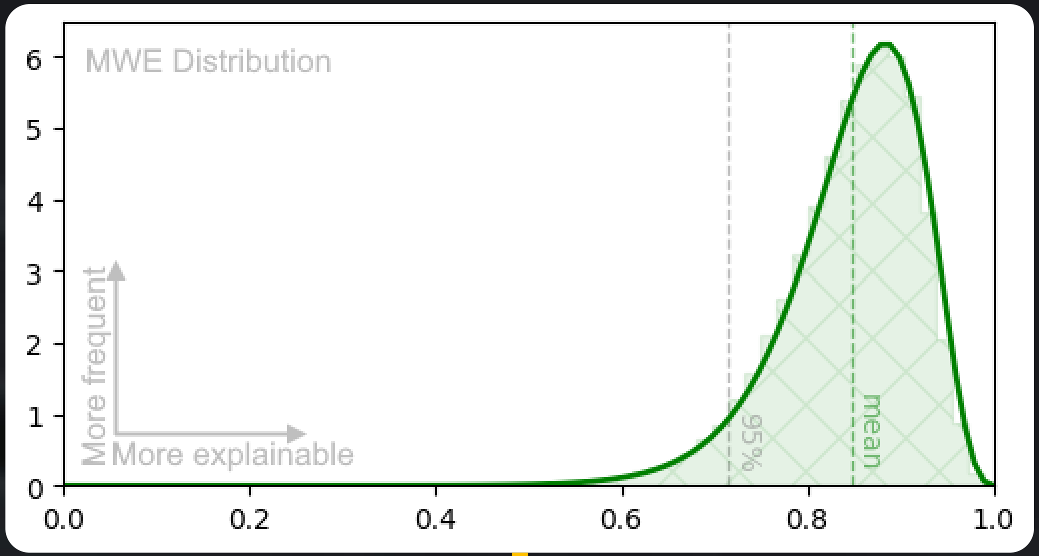


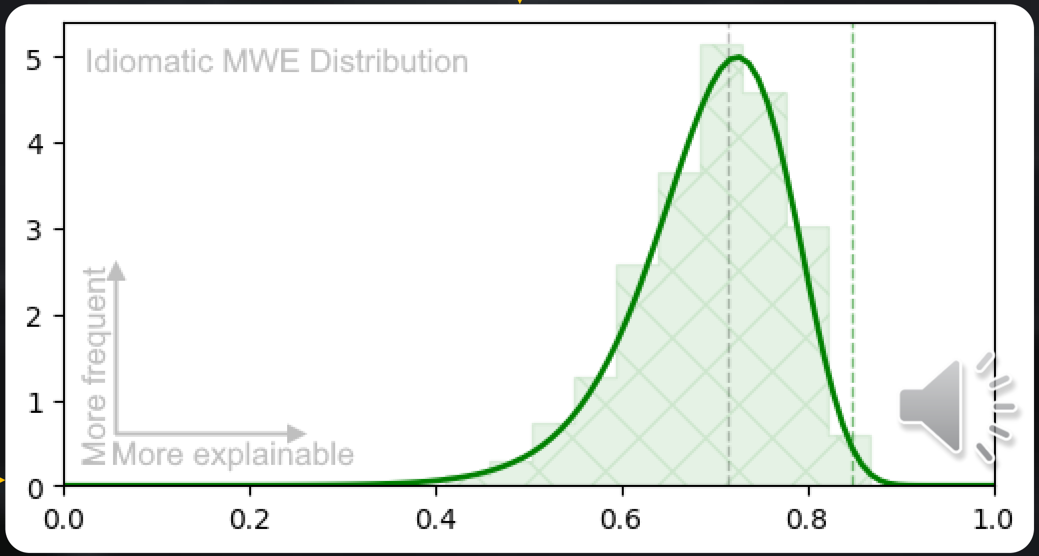
RESULTS

We sample the MWEs by score bucket, and classify them between self-explanatory and idiomatic…

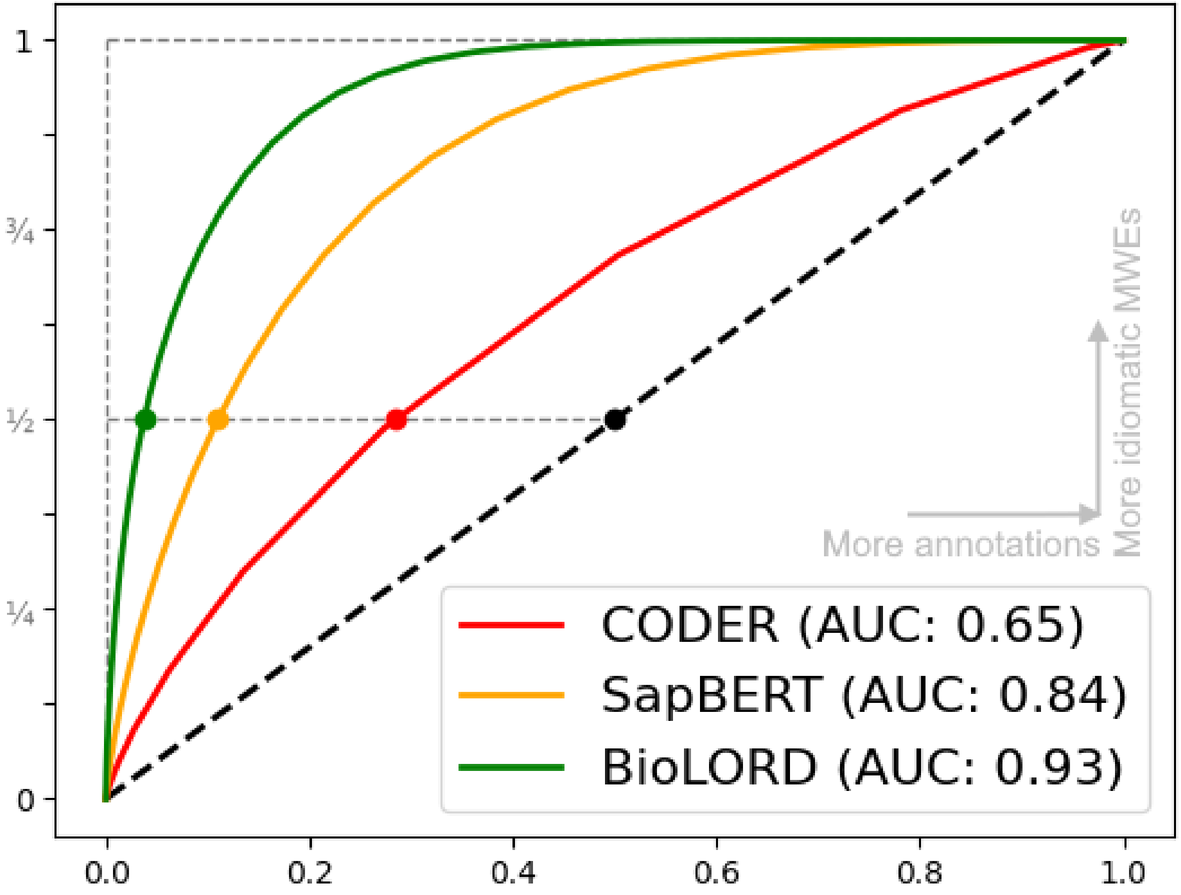
ANNOTATIONS







RESULTS (ROC CURVE)



To recall 50% of all idiomatic expressions in the dataset, it is only required to annotate 5% of the data using the BioLORD-based score.

Other models, which do not ground their representation using definitions, perform significantly worse.

THE END

Thank you

for listening to this talk

@FremyCompany

francois.remy@ugent.be