A Corpus of Biology Analogy Questions as a Challenge for Explainable AI

Han Lin Aung, Justin Xu, Sajana Weerawardhena, Vinay Chaudhri

A is to B as C is to what?

Man: King as Woman: Queen

Motivation

Analogy as a learning tool for students in discovering relations esp. in science

Intelligent Biology Textbook (Inquire)

Related work on analogy datasets

Google Analogy Dataset[1]

The Bigger Analogy Test Set (BATS)[2]

[1] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In Proceedings of International Conference on Learning Representations (ICLR).

[2] Gladkova, A., Drozd, A., & Matsuoka, S. (2016). Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn’t. In Proceedings of the NAACL-HLT SRW (pp. 47–54). San Diego, California, June 12-17, 2016: ACL. Retrieved from https://www.aclweb.org/anthology/N/N16/N16-2002.pdf

Related Work on answering analogy questions

Word embedding[1]

SemEval 2012 Task 2[2]

- Measuring Degrees of Relational Similarity

Pair-pattern matrix[3]

Deep learning

- BERT[4]

1. Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In EMNLP, 2014
2. David A. Jurgens, Saif M. Mohammad, Peter D. Turney, and Keith J. Holyoak (2012), SemEval-2012 Task 2: Measuring Degrees of Relational Similarity, First Joint Conference on Lexical and Computational Semantics (*SEM), Montreal, Canada, June 2012, pp. 356–364.
3. Peter D. Turney and Patrick Pantel. From frequency to meaning: Vector space models of semantics.J. Artif. Int. Res., 37(1):141–188, January 2010.
4. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018

Dataset

Biology Analogy Questions Corpus:

Analogy questions generated from Biology Knowledge Base

Text data:

LIFE Biology textbook

OpenStax Biology textbooks

Generation of biology analogy corpus using knowledge base

Generation of Biology Analogy Questions

KB_Bio_101[1]: hand-curated knowledge base

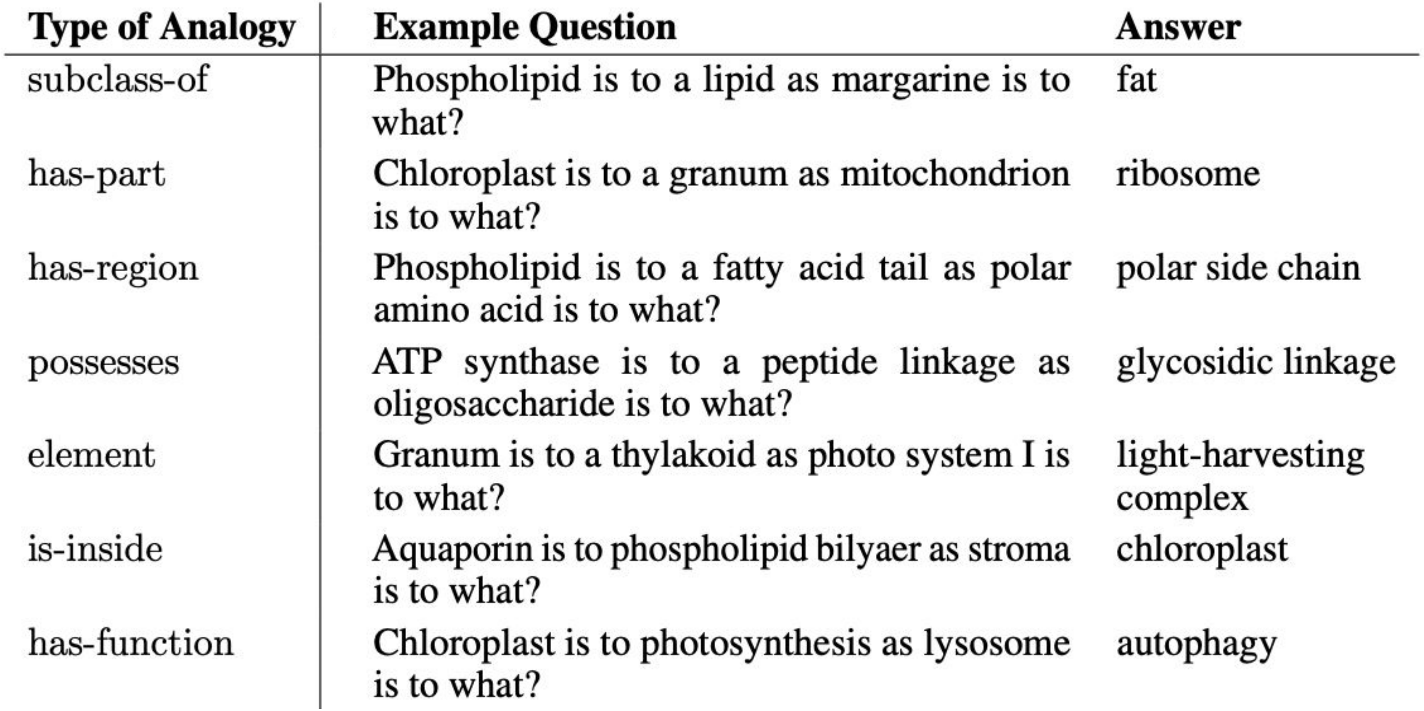
Extraction of analogy from knowledge base

- Crawl of knowledge base of different semantic relationships in **KB_Bio_101**
- Relations include: subclass-of, has-part, has-region, possesses, etc
- Allows interpretability of **multi-hop** analogies
  - E.g. **A** SUBCLASS-OF **B** SUBCLASS-OF **C**
- 70,000+ analogy questions

[1] Vinay K. Chaudhri, Daniel Elenius, Sue Hinojoza, Michael A. Wessel. KB_Bio_101: Content and Challenges. In *In the Proceedings of International Conference on Formal Ontologies in Information Systems, 2014.*

Generation of Biology Analogy Questions

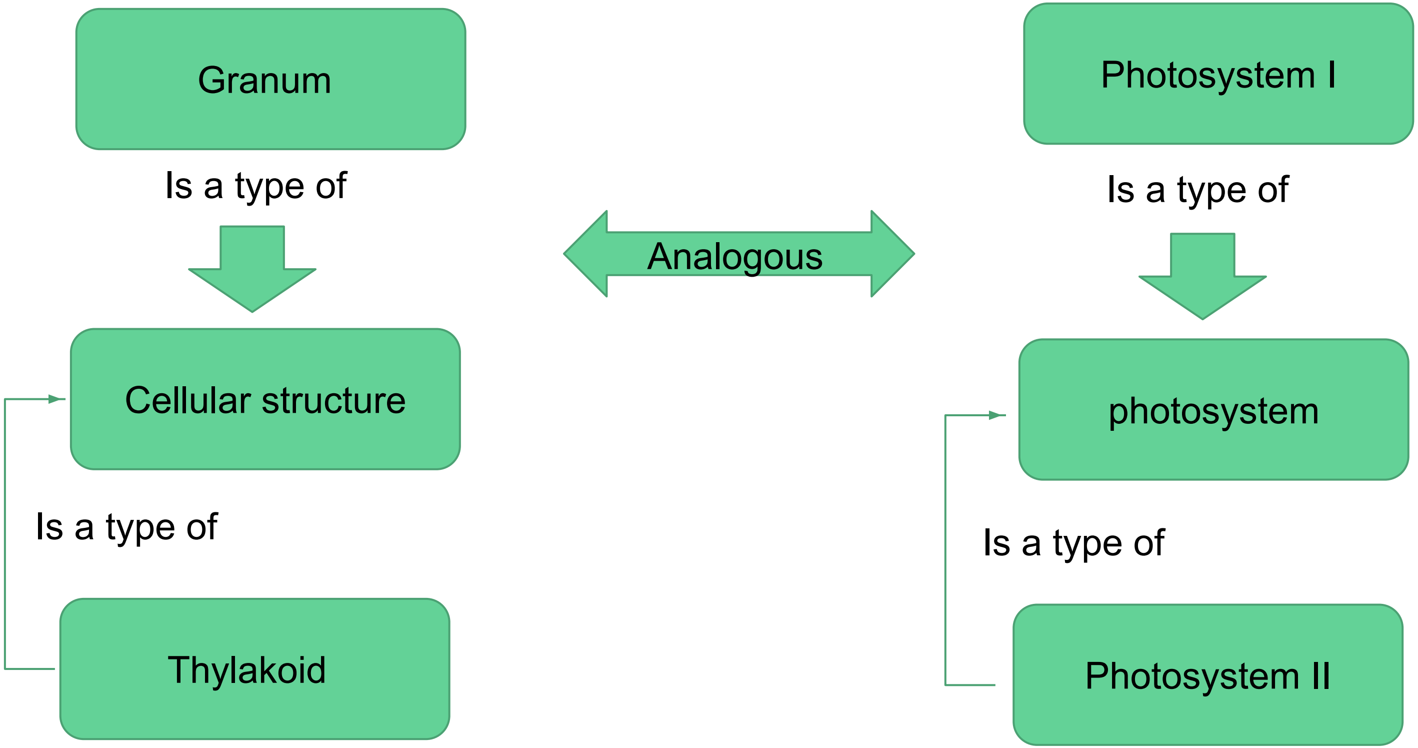
Analogy examples



Type of Analogy	Example Question	Answer
subclass-of	Phospholipid is to a lipid as margarine is to what?	fat
has-part	Chloroplast is to a granum as mitochondrion is to what?	ribosome
has-region	Phospholipid is to a fatty acid tail as polar amino acid is to what?	polar side chain
possesses	ATP synthase is to a peptide linkage as oligosaccharide is to what?	glycosidic linkage
element	Granum is to a thylakoid as photo system I is to what?	light-harvesting complex
is-inside	Aquaporin is to phospholipid bilyaer as stroma is to what?	chloroplast
has-function	Chloroplast is to photosynthesis as lysosome is to what?	autophagy

Generation of Biology Analogy Questions

Interpretability


Granum

Is a type of

Cellular structure

Is a type of

Thylakoid

Analogous

Photosystem I

Is a type of

photosystem

Is a type of

Photosystem II

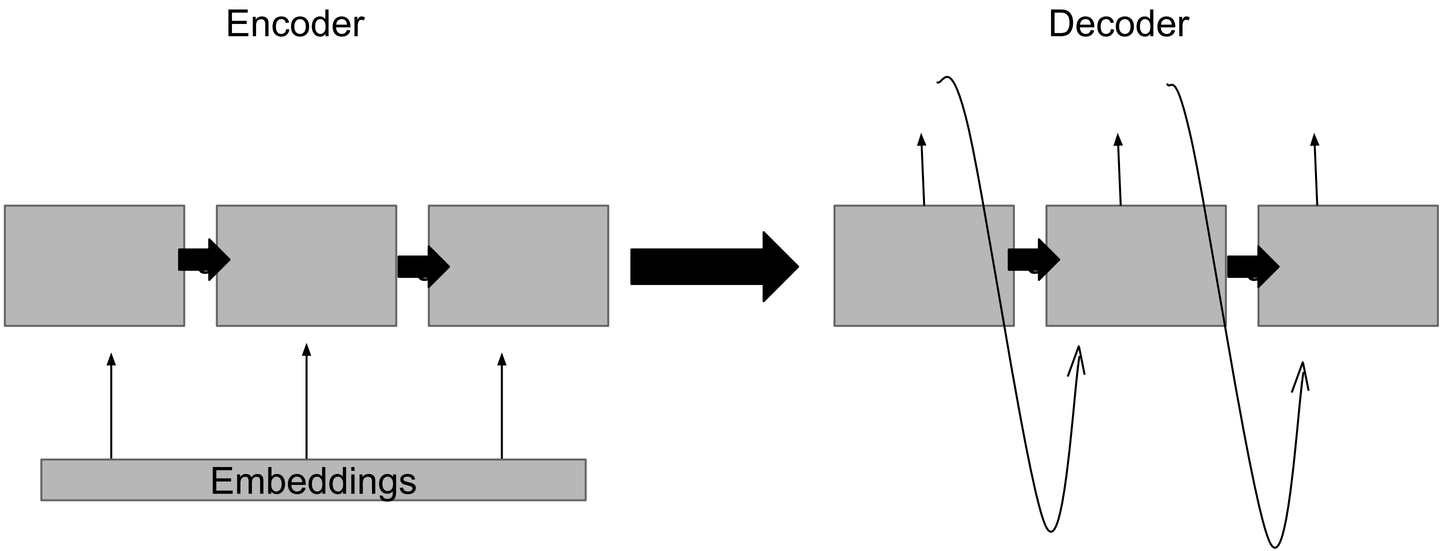
Automated answering of analogy questions

Automated answering of analogy questions

	GLoVe embeddings	FastText embeddings	ELMo embeddings	BERT/BioBERT embeddings
Word embeddings	x	x	x	
Seq2Seq	x		x	
Seq2Vec	x		x	x

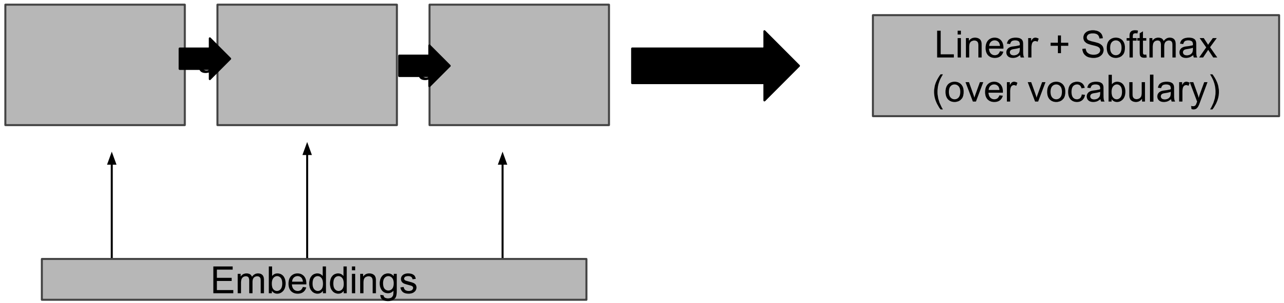
The word embeddings are trained on the LIFE Biology and OpenStax textbooks.

Seq2Seq



▨ Each block is a LSTM cell. The encoder is bi-directional LSTM.

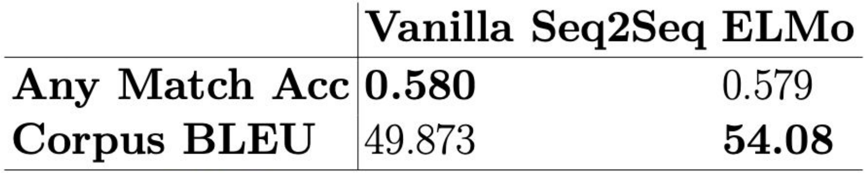
Seq2Vec



▨ Each block is a LSTM cell. The encoder is bi-directional LSTM.

Results

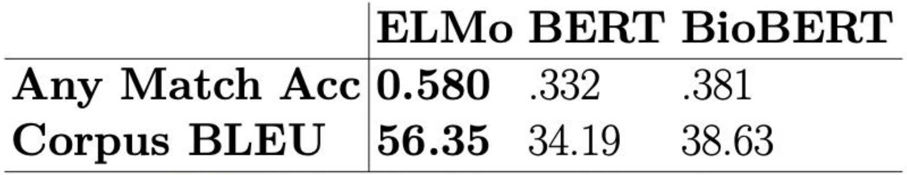
Seq2Seq



	Vanilla Seq2Seq	ELMo
Any Match Acc	**0.580**	0.579
Corpus BLEU	49.873	**54.08**

Table 2: Best Seq2Seq models results

Seq2Vec



	ELMo	BERT	BioBERT
Any Match Acc	**0.580**	.332	.381
Corpus BLEU	**56.35**	34.19	38.63

Table 3: Best Seq2Vec models results

Results - Examples (Seq2Vec)

Correct examples:

Nonpolar covalent bond | chemical bond | ultraviolet ray : **light**

- Prediction: light

carbon atom | atom | cytochrome A3 : **cytochrome**

- Prediction: cytochrome

Incorrect example:

thymine | pyrimidine | water molecule : **polar molecule** (Wrong relationship)

Prediction: **oxygen atom**

Conclusion & Future Work

Utility of knowledge base

- Interpretability, generation of questions

Automated interpretability of analogies

- Relation extraction as a means for explanation

Open problem for more research!

Comments, Questions, Suggestions?

Appendix: More info on dataset

Data split on corpus:

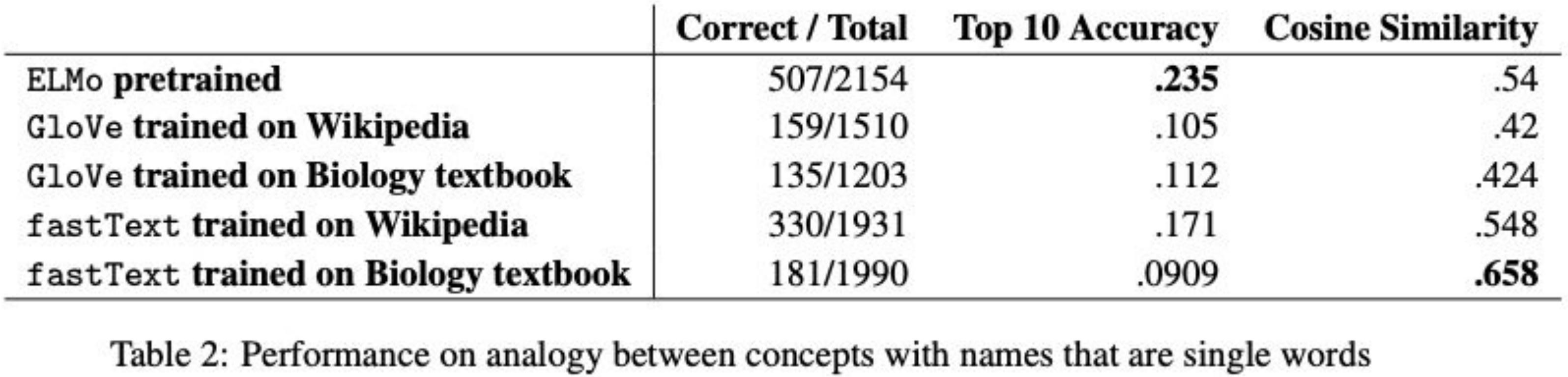
Train/val/test: 70/10/20

Specific Future directions for analogy evaluation:

- Currently treats multi-hop relations that have same relations for all hops as one relation in dataset evaluation (causes some analogies to be “far-fetched”)
- Quantitative Evaluation on analogy by domain expert
- Evaluation on unseen concepts for analogies

Appendix: Word embedding evaluation

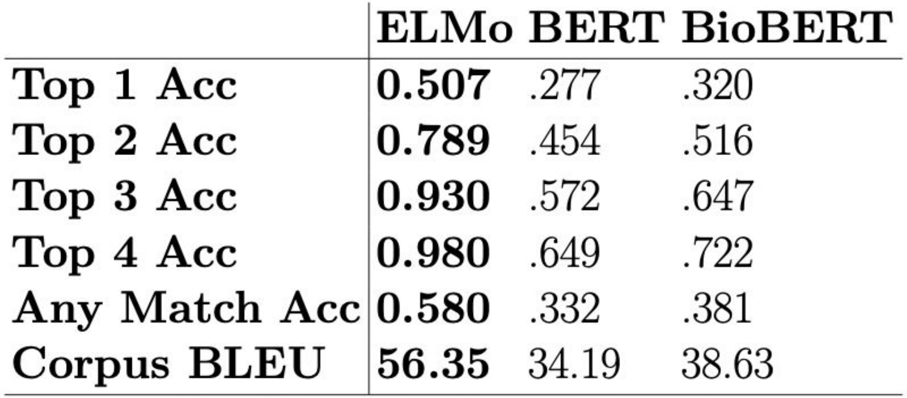
Word embeddings evaluation



	Correct / Total	**Top 10 Accuracy**	**Cosine Similarity**
`ELMo` **pretrained**	507/2154	**.235**	.54
`GloVe` **trained on Wikipedia**	159/1510	.105	.42
`GloVe` **trained on Biology textbook**	135/1203	.112	.424
`fastText` **trained on Wikipedia**	330/1931	.171	.548
`fastText` **trained on Biology textbook**	181/1990	.0909	**.658**

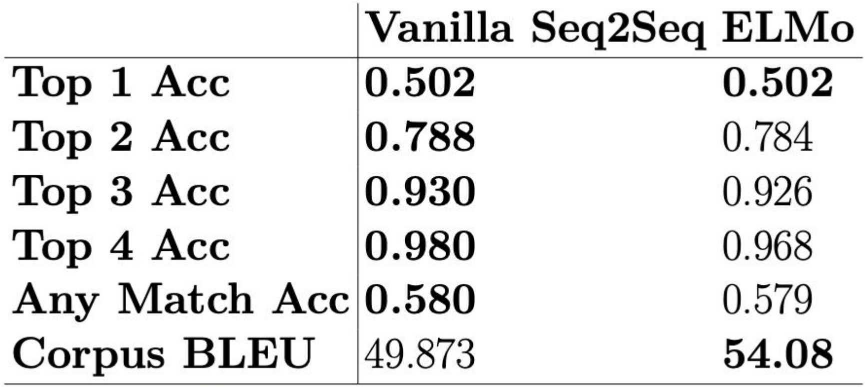
Table 2: Performance on analogy between concepts with names that are single words

Appendix: Specific results



	ELMo	**BERT**	**BioBERT**
Top 1 Acc	**0.507**	.277	.320
Top 2 Acc	**0.789**	.454	.516
Top 3 Acc	**0.930**	.572	.647
Top 4 Acc	**0.980**	.649	.722
Any Match Acc	**0.580**	.332	.381
Corpus BLEU	**56.35**	34.19	38.63

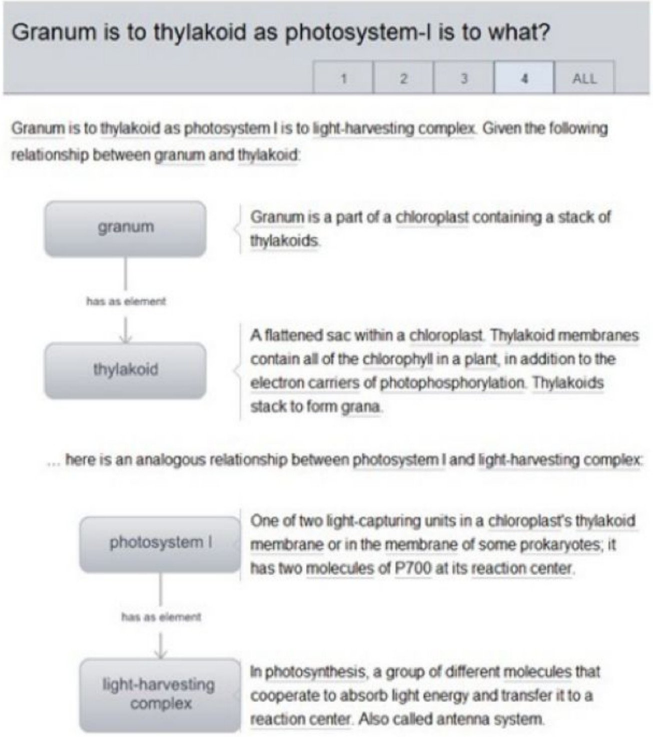
Table 3: Best Seq2Vec models results



	Vanilla Seq2Seq	**ELMo**
Top 1 Acc	**0.502**	**0.502**
Top 2 Acc	**0.788**	0.784
Top 3 Acc	**0.930**	0.926
Top 4 Acc	**0.980**	0.968
Any Match Acc	**0.580**	0.579
Corpus BLEU	49.873	**54.08**

Table 2: Best Seq2Seq models results

Appendix: Specific interpretability results



Granum is to thylakoid as photosystem-I is to what?

Granum is to thylakoid as photosystem I is to light-harvesting complex. Given the following relationship between granum and thylakoid:

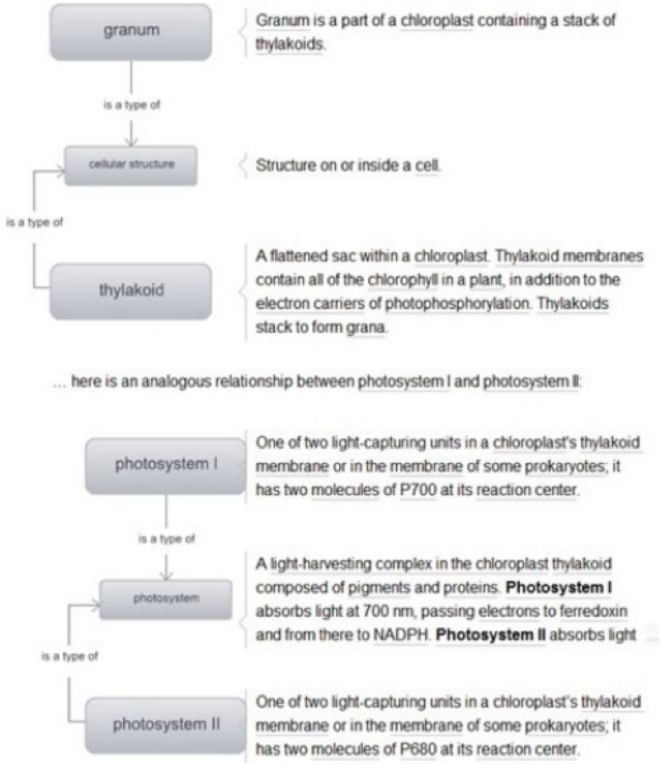
Granum is a part of a chloroplast containing a stack of thylakoids.

A flattened sac within a chloroplast. Thylakoid membranes contain all of the chlorophyll in a plant, in addition to the electron carriers of photophosphorylation. Thylakoids stack to form grana.

... here is an analogous relationship between photosystem I and light-harvesting complex:

One of two light-capturing units in a chloroplast's thylakoid membrane or in the membrane of some prokaryotes; it has two molecules of P700 at its reaction center.

In photosynthesis, a group of different molecules that cooperate to absorb light energy and transfer it to a reaction center. Also called antenna system.



Granum is a part of a chloroplast containing a stack of thylakoids.

Structure on or inside a cell.

A flattened sac within a chloroplast. Thylakoid membranes contain all of the chlorophyll in a plant, in addition to the electron carriers of photophosphorylation. Thylakoids stack to form grana.

... here is an analogous relationship between photosystem I and photosystem II:

One of two light-capturing units in a chloroplast's thylakoid membrane or in the membrane of some prokaryotes; it has two molecules of P700 at its reaction center.

A light-harvesting complex in the chloroplast thylakoid composed of pigments and proteins. **Photosystem I** absorbs light at 700 nm, passing electrons to ferredoxin and from there to NADPH. **Photosystem II** absorbs light ...

One of two light-capturing units in a chloroplast's thylakoid membrane or in the membrane of some prokaryotes; it has two molecules of P680 at its reaction center.

Appendix: Additional examples

More correct examples

diacylglycerol | amphipathic molecule | nucleotide | molecule

Prediction: organic molecule

AMP | phosphate group | AMP | carbon skeleton

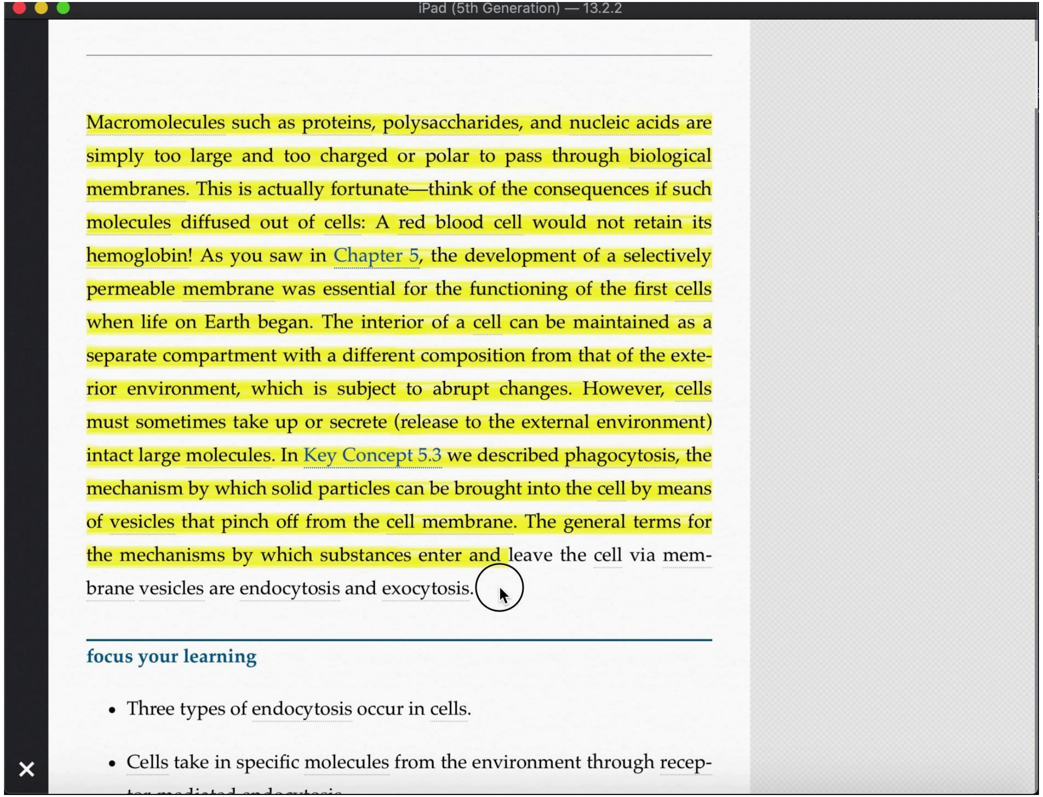
Prediction: carbon skeleton

“Incorrect” examples from data:

glucose | polar covalent bond | oxygen molecule | nonpolar covalent bond
Prediction: double bond

nonpolar covalent bond | covalent bond | plant cell-wall | cell wall
Prediction: cellular structure

Appendix: Full demo



Macromolecules such as proteins, polysaccharides, and nucleic acids are simply too large and too charged or polar to pass through biological membranes. This is actually fortunate—think of the consequences if such molecules diffused out of cells: A red blood cell would not retain its hemoglobin! As you saw in Chapter 5, the development of a selectively permeable membrane was essential for the functioning of the first cells when life on Earth began. The interior of a cell can be maintained as a separate compartment with a different composition from that of the exterior environment, which is subject to abrupt changes. However, cells must sometimes take up or secrete (release to the external environment) intact large molecules. In Key Concept 5.3 we described phagocytosis, the mechanism by which solid particles can be brought into the cell by means of vesicles that pinch off from the cell membrane. The general terms for the mechanisms by which substances enter and leave the cell via membrane vesicles are endocytosis and exocytosis.

focus your learning

- Three types of endocytosis occur in cells.
- Cells take in specific molecules from the environment through receptor-mediated endocytosis.