GPT4 aided biomaterials research use case : stabilization of selenium nanoparticles with proteins

Zoran Stojanović,* Nenad Filipović, Maja Kuzmanović, Sara Lukač, Magdalena Stevanović

Institut tehničkih nauka SANU

Institute of Technical Sciences of SASA, Belgrade, Republic of Serbia

Introduction

Recent advancements in LLMs based on various transformer architectures such as BERT and GPT family models, brought many new possibilities for application in scientific research. The specific architecture and broad knowledge of these models give them the ability to understand concepts, to plan and solve different kinds of problems, including various chemistry – related tasks.

We evaluated a case of GPT4 performance for recommending proteins suitable for the stabilization of selenium nanoparticles (SeNPs). SeNPs exhibit diverse beneficial bioactivities, including antioxidant, antibacterial, and anticancer properties.

Stabilization of SeNPs with suitable proteins may be an effective approach to improve their bioactivities.

34 Se Selenium 78.971

Methods

- **Prompt engineering:**
  - GPT4 - turbo - preview
  - Open AI playground and API
  - Python 3.10 virtual environment
- **Data:**
  - Science parse AllenAI
    - Research articles PDFs
    - Review articles PDFs
  - UniProt API for protein records
- **Chemistry knowledge assessment:**
  - Chemistry knowledge and reasoning
  - SeNp synthesis and stabilization methods
  - Synthesis text classification
  - Information extraction to tabular data – JSON format
- **Protein knowledge:**
  - Properties and functions of proteins
  - Amino acid sequence recognition
  - Comparing amino acid sequences (failed)
  - Propose proteins with desired bioactivites after examples provided

github.com/zorankiki/gpt4_for_SeNp_research

uniprot.org/help/api

Classification prompt example

Extraction prompt example

BSA UniProt ID Test

Conclusion

The study demonstrates the successful application of advanced transformer architecture models like GPT4 in addressing relatively complex tasks in materials research. Despite GPT4 capabilities being largely dependent on the quality and size of training data, utilization of strategically designed and optimized prompts significantly improves it's performance in many cases. Although models can not perform well some "trivial" tasks, such as find longest common substring between two (or more) protein sequences, it has a great potential in research design and planing.

Results

- **Chemistry knowledge assessment**
  - Classification task accuracy 96%
  - Synthesis information extraction task

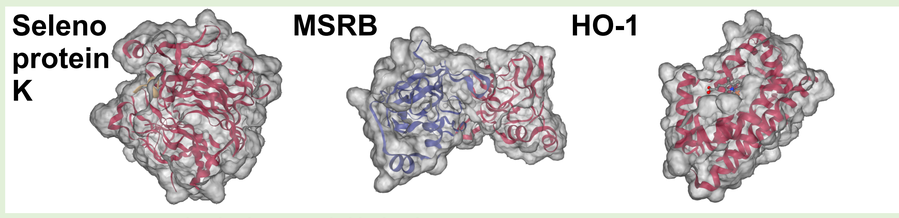
	pos	neg
neg	9	275
pos	10	2

title	synthesis_approach	stabilizing_agent
45S5Bioglass®-based scaffolds coated with sele...	[foam replica method, physicochemical solvent/...	[bovine serum albumin (BSA)]
Bovine Serum Albumin Nanospheres Synchronously...	[Spray drying]	[BSA]
Bovine Serum Albumin Nanospheres Synchronously...	[]	[BSA]
Comparative Study of the Antimicrobial Activit...	[chemical reduction]	[BSA, Chitosan]
Mediated by tea polypeptides: A green synthesi...	[green synthesis]	[tea protein]
Poly (ε-caprolactone) microspheres for prolong...	[chemical reduction]	[BSA]

- Protein stabilizers extraction task from reviews – proteins that interacts with Se

molecule	freq	desc	molecule	freq	desc	molecule	freq	desc	molecule	freq	desc	molecule	freq	desc
bsa	4	None	glutathione peroxidase	2	None	sparc	1	secreted protein and rich in cysteine	selenocysteine	1	None	cytochrome c4	1	None
lysozyme	3	None	insulin	2	None	pacap-derived peptide dbayl	1	pituitary adenylate catalase activating peptide	selenos	1	None	dio2	1	selenoprotein
transferrin	3	None	chitosan	2	polysacharide	eadf4(k16)	1	positively charged spider silk protein eADF4(k16)	reductase protein	1	None	selenon	1	None
mushroom polysaccharide protein complex	2	None	selenok	1	None	horseradish peroxidase	1	None	polysaccharide and protein complex of edible m...	1	None	selenof	1	None
human serum albumin	2	None	keratin	1	None	hsp-70	1	heat shock protein	thioredoxin reductase	1	None	selenot	1	None
silk fibroin	2	None	streptavidin	1	None	psp	1	polysaccharide–protein	selenomethionine	1	None	selenom	1	None
cytochrome c3	2	None	ox26	1	monoclonal antibodies antitransferin recepotor	selenoproteins	1	None	sef a	1	Sef A protein	rgdfc	1	RGDfC (cyclic peptide)

- Examples of recommended proteins by model

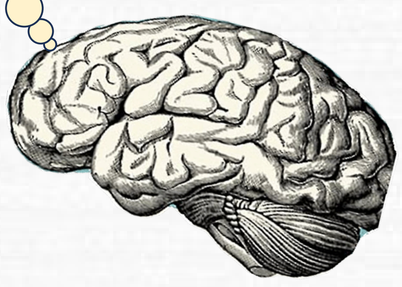


 # who you are: you are research scientist, expert in biochemistry.
 # your task: propose human proteins that interact with Selenium (Se) which have: a) anti-inflammatory; b) anti-apoptotic and c) anti-cancer role. You will get data of human proteins known to interact with Se. Propose proteins that are not listed in data that have stated roles (a, b, c, or combination of two or all together).
 # data format example: {'uniprot_name': {0: 'protein_0', 1: 'protein_1', ...}, {'uniprot_id': {0: 'ID_protein_0', 1: 'ID_protein_1', ...}, 'seq': {0: 'seq_protein_0', 1: 'seq_protein_1', ...}, ...}
 # data fields and explanation: 'uniprot_name' - name of protein in UniProt DB; 'uniprot_id' - id in UniProt DB; 'seq' - amino acid sequence of protein, 'function' - description of protein function in human body, 'simmilarity' - protein family, 'ctn_by_feature_type' - protein features like binding sites, helices , etc; 'resource_url' - external url usually Wikipedia entry.

- Human proteins stabilizers examples data to prompt

uniprot_name	uniprot_id	seq	function	simmilarity	ctn_by_feature_type	interactions	resource_url
Lysozyme C	P61626	MKALIVLGLVLLSVTVQGKVFERCELARTLKRLGMDGYRGISLANW...	Lysozymes have primarily a bacteriolytic funct...	Belongs to the glycosyl hydrolase 22 family	{'Signal': 1, 'Chain': 1, 'Domain': 1, 'Active...	[[{'interactantOne': {'uniProtKBAccession': 'P...	https://en.wikipedia.org/wiki/Lysozyme
Serotransferrin	P02787	MRLAVGALLVCAVLGLCLAVPDKTVRWCAVSEHEATKCQSFRDHMK...	Transferrins are iron binding transport protei...	Belongs to the transferrin family	{'Signal': 1, 'Chain': 1, 'Domain': 2, 'Bindin...	[[{'interactantOne': {'uniProtKBAccession': 'P...	https://en.wikipedia.org/wiki/Transferrin
Albumin	P02768	MKWVTFISLLFLFSSAYSRGVFRRDAHKSEVAHRFKDLGEENFKAL...	Binds water, Ca(2+), Na(+), K(+), fatty acids,...	Belongs to the ALB/AFP/VDB family	{'Signal': 1, 'Propeptide': 1, 'Chain': 1, 'Do...	[[{'interactantOne': {'uniProtKBAccession': 'P...	https://albumin.org
Glutathione peroxidase 1	P07203	MCAARLAAAAAAAQSVYAFSARPLAGGEPVSLGSLRGKVLLIENVA...	Catalyzes the reduction of hydroperoxides in a...	Belongs to the glutathione peroxidase family	{'Chain': 1, 'Active site': 1, 'Site': 1, 'Non...	[[{'interactantOne': {'uniProtKBAccession': 'P...	http://egp.gs.washington.edu/data/gpx1/
Insulin	P01308	MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGER...	Insulin decreases blood glucose concentration....	Belongs to the insulin family	{'Signal': 1, 'Peptide': 2, 'Propeptide': 1, '...	[[{'interactantOne': {'uniProtKBAccession': 'P...	https://www.lillydiabetes.com/assets/pdf/pp-id...
Keratin, type II cytoskeletal 2 epidermal	P35908	MSCQISCKSRGRGGGGGGFRGFSSGSAVVSGGSRRSTSSFSCLSRH...	Probably contributes to terminal cornification...	Belongs to the intermediate filament family	{'Chain': 1, 'Domain': 1, 'Region': 9, 'Compos...	[[{'interactantOne': {'uniProtKBAccession': 'P...	https://en.wikipedia.org/wiki/Keratin_2A
SPARC	P09486	MRAWIFFLLCLAGRALAAPQQEALPDETEVVEETVAEVTEVSVGAN...	Appears to regulate cell growth through intera...	Belongs to the SPARC family	{'Signal': 1, 'Chain': 1, 'Domain': 3, 'Bindin...	[[{'interactantOne': {'uniProtKBAccession': 'P...	https://en.wikipedia.org/wiki/Osteonectin
Flavin reductase (NADPH)	P30043	MAVKKIAIFGATGQTGLTTLAQAVQAGYEVTVLVRDSSRLPSEGPR...	Broad specificity oxidoreductase that catalyze...	NaN	{'Initiator methionine': 1, 'Chain': 1, 'Bindi...	[[{'interactantOne': {'uniProtKBAccession': 'P...	http://egp.gs.washington.edu/data/blvrb/
Thioredoxin reductase 1, cytoplasmic	Q16881	MGCAEGKAVAAAAPTELQTKGKNGDGRRRSAKDHHPGKTLPENPAG...	Reduces disulfideprotein thioredoxin (Trx) to ...	Belongs to the class-I pyridine nucleotide-dis...	{'Chain': 1, 'Domain': 1, 'Region': 2, 'Compos...	[[{'interactantOne': {'uniProtKBAccession': 'Q...	http://egp.gs.washington.edu/data/txnrd1/

LLMs, what was that?



e-mail: zoran.stojanovic@itn.sanu.ac.rs

LinkedIn: https://www.linkedin.com/in/drzoki/

ORCID: https://orcid.org/0000-0002-5989-0031

ResearchGate: https://www.researchgate.net/profile/Zoran-Stojanovic

Acknowledgment

This work was funded by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia, grant 451-03-66/2024-03/200175