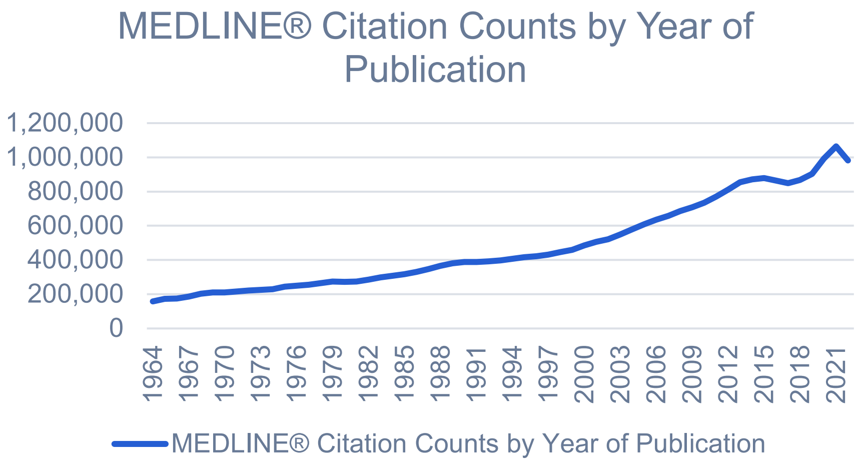
SciMON 🧪 : Scientific Inspiration Machines Optimized for Novelty

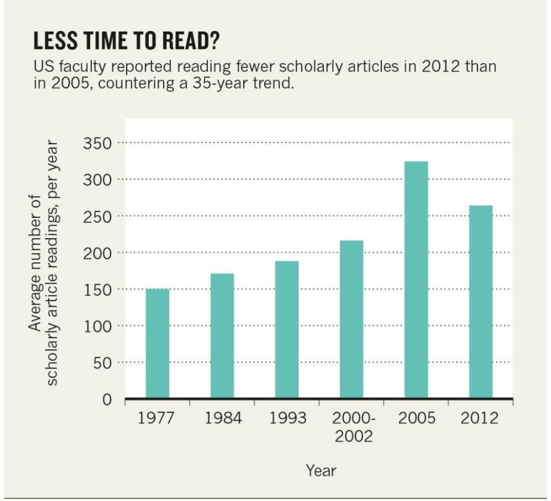
Qingyun Wang[1], Doug Downey[2], Heng Ji[1], Tom Hope[2,3]

[1]University of Illinois at Urbana-Champaign
[2]Allen Institute for Artificial Intelligence (AI2)
[3]The Hebrew University of Jerusalem

Motivation: Augmenting Human Innovation

- Millions of scientific papers are published every year
- Human’s reading ability keeps almost the same across years



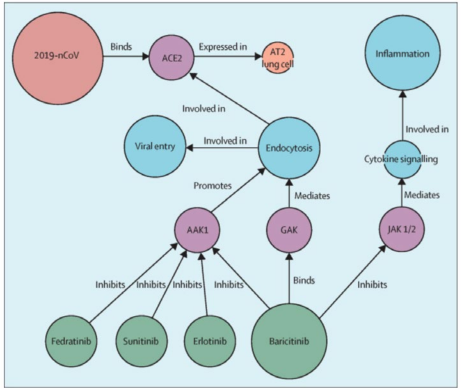

LESS TIME TO READ?

US faculty reported reading fewer scholarly articles in 2012 than in 2005, countering a 35-year trend.

https://www.nlm.nih.gov/bsd/medline_cit_counts_yr_pub.html
Landhuis, E. (2016). Scientific literature: Information overload. Nature, 535(7612), 457-458

Background: Scientific Knowledge Discovery

- Literature-based Discovery
  - Limited to curated entities and relations
  - Limited to certain domains
  - Cannot model nuanced contexts
- LLMs for Scientific Innovation
  - Limited to code generation/experiment planning
  - Focusing on anecdotal evaluation



THE NATIONAL CENTER FOR BIOMEDICAL ONTOLOGY

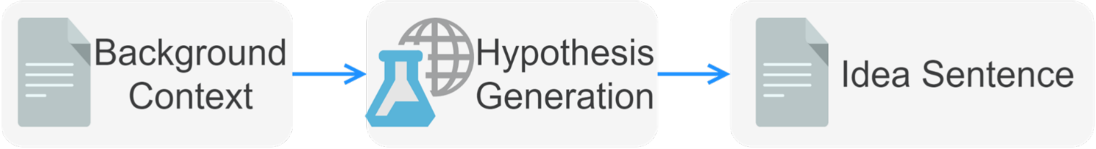
Ontobee

NIH NATIONAL CANCER INSTITUTE Center for Biomedical Informatics & Information Technology

BioPortal

LIMICS

Contextualized Literature-based Discovery



Seed Term: **knowledge acquisition**
Background: ... This requires plms to **integrate the information from all the sources** in a lifelong manner. Although this goal could be achieved by exhaustive pre-training on all the existing data, such a process is known to be **computationally expensive.**

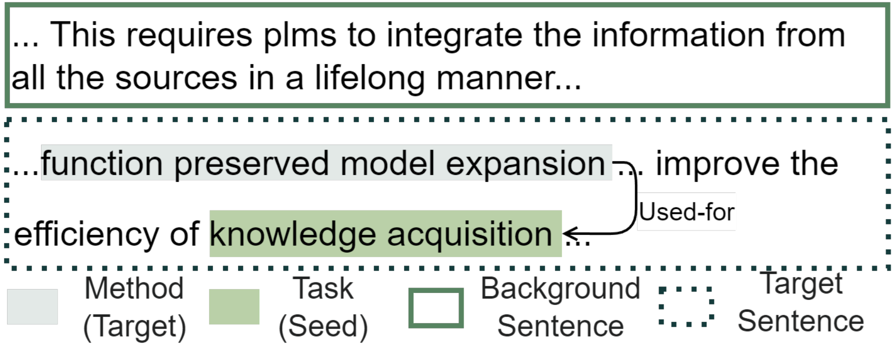
Specifically, ELLE consists of (1) **function preserved model expansion**, which flexibly expands an existing PLM's width and depth to improve the efficiency of **knowledge acquisition** ...

Qin, Y., Zhang, J., Lin, Y., Liu, Z., Li, P., Sun, M., & Zhou, J. (2022). ELLE: Efficient lifelong pre-training for emerging data. ACL Findings 2022.

Dataset Construction

- Construct a corpus from 67,408 ACL Anthology papers from 1952 to 2022 with 5,946 papers from 2021, and 2,588 papers from 2022
- Focus on *used-for* relations, which usually include tasks and methods

Split	Forward	Backward	Total
Train	55,884	58,426	114,310
Valid	7,938	8,257	16,195
Test	2,623	2,686	5,309

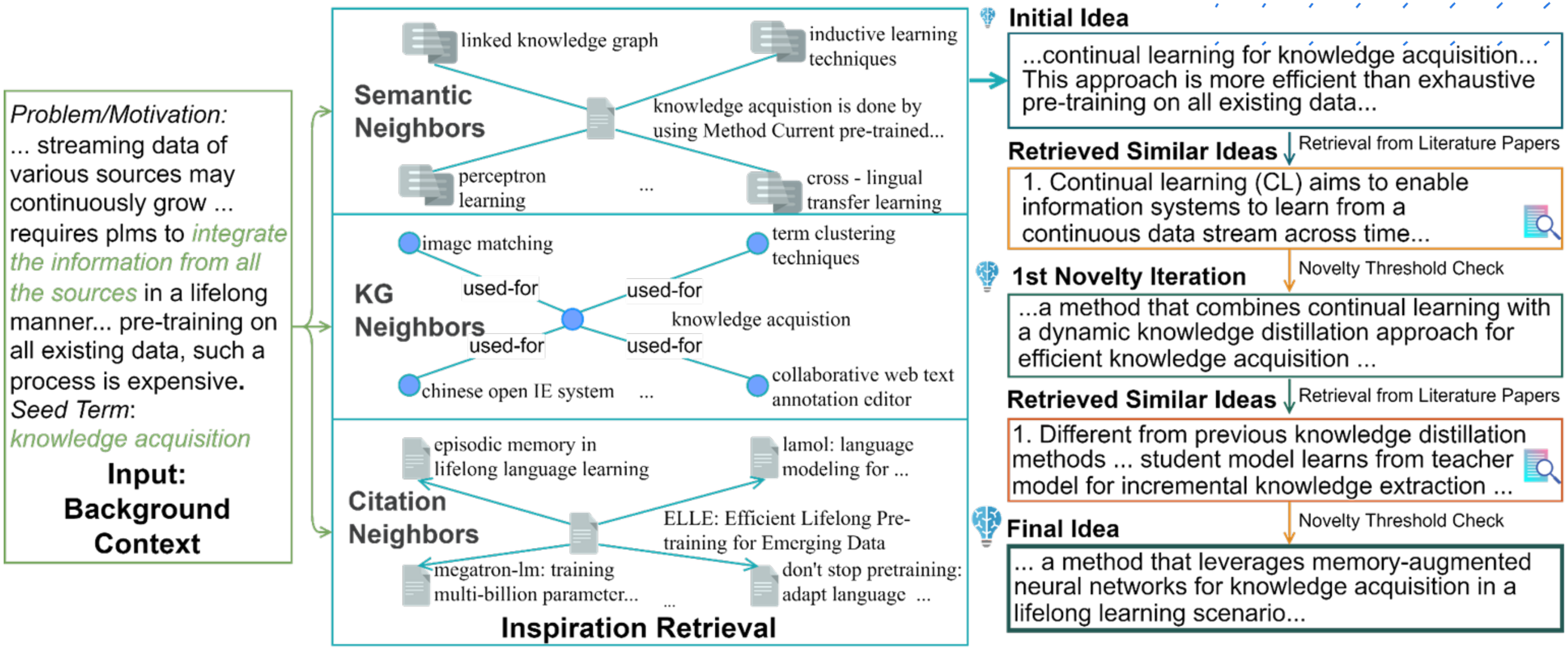


... This requires plms to integrate the information from all the sources in a lifelong manner...

...function preserved model expansion ... improve the efficiency of knowledge acquisition ...

Qin, Y., Zhang, J., Lin, Y., Liu, Z., Li, P., Sun, M., & Zhou, J. (2022). ELLE: Efficient lifelong pre-training for emerging data. ACL Findings 2022.

SciMON Overview



Problem/Motivation: ... streaming data of various sources may continuously grow ... requires plms to *integrate the information from all the sources* in a lifelong manner... pre-training on all existing data, such a process is expensive.

Seed Term: *knowledge acquisition*

Input: Background Context

Semantic Neighbors
linked knowledge graph
inductive learning techniques
knowledge acquistion is done by using Method Current pre-trained...
perceptron learning
...
cross - lingual transfer learning
KG Neighbors
image matching
term clustering techniques
used-for
used-for
knowledge acquistion
used-for
used-for
chinese open IE system
...
collaborative web text annotation editor
Citation Neighbors
episodic memory in lifelong language learning
lamol: language modeling for ...
ELLE: Efficient Lifelong Pre-training for Emerging Data
megatron-lm: training multi-billion parameter...
...
don't stop pretraining: adapt language ...
Inspiration Retrieval

Initial Idea

...continual learning for knowledge acquisition... This approach is more efficient than exhaustive pre-training on all existing data...

Retrieved Similar Ideas Retrieval from Literature Papers

1. Continual learning (CL) aims to enable information systems to learn from a continuous data stream across time...

1st Novelty Iteration Novelty Threshold Check

...a method that combines continual learning with a dynamic knowledge distillation approach for efficient knowledge acquisition ...

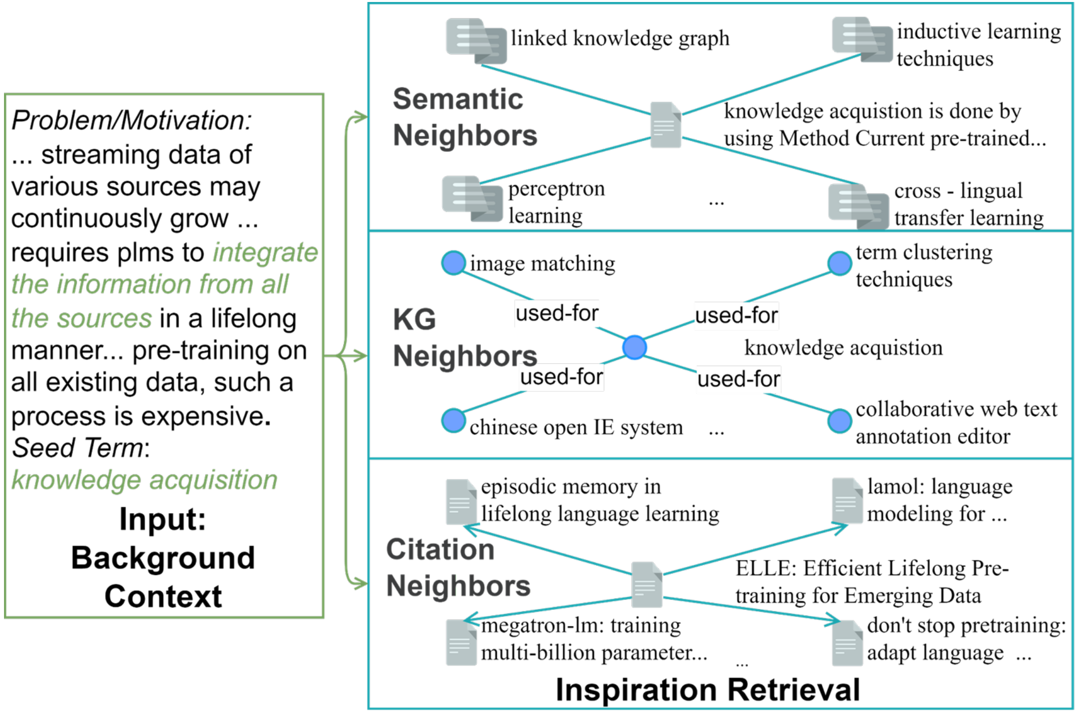
Retrieved Similar Ideas Retrieval from Literature Papers

1. Different from previous knowledge distillation methods ... student model learns from teacher model for incremental knowledge extraction ...

Final Idea Novelty Threshold Check

... a method that leverages memory-augmented neural networks for knowledge acquisition in a lifelong learning scenario...

Inspiration Retrieval



Problem/Motivation: ... streaming data of various sources may continuously grow ... requires plms to *integrate the information from all the sources* in a lifelong manner... pre-training on all existing data, such a process is expensive.

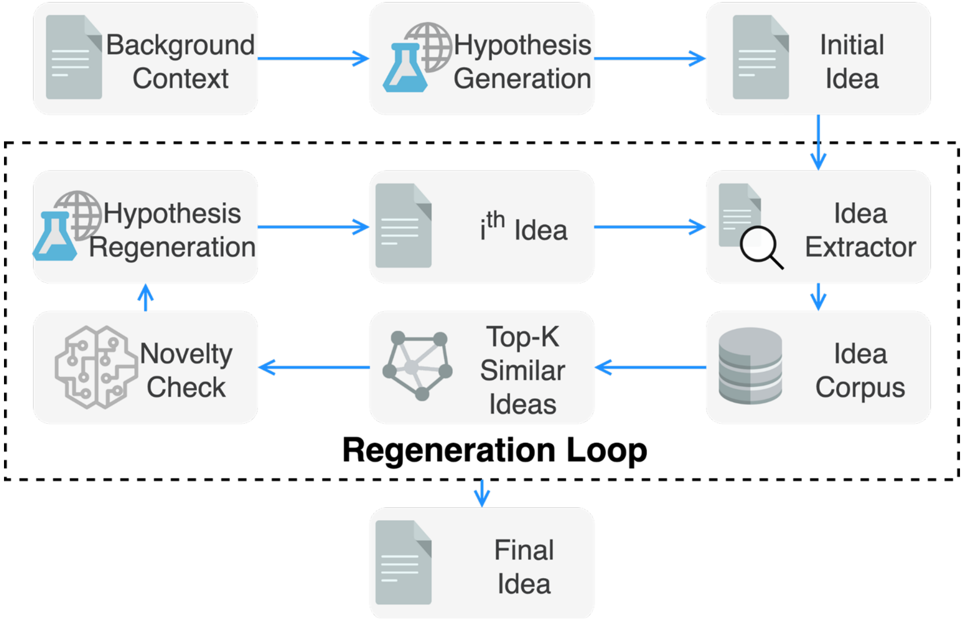
Seed Term: *knowledge acquisition*

Input: Background Context

Semantic Neighbors
linked knowledge graph
inductive learning techniques
knowledge acquistion is done by using Method Current pre-trained...
perceptron learning
...
cross - lingual transfer learning
KG Neighbors
image matching
term clustering techniques
used-for
used-for
knowledge acquistion
used-for
used-for
chinese open IE system
...
collaborative web text annotation editor
Citation Neighbors
episodic memory in lifelong language learning
lamol: language modeling for ...
ELLE: Efficient Lifelong Pre-training for Emerging Data
megatron-lm: training multi-billion parameter...
...
don't stop pretraining: adapt language ...
Inspiration Retrieval

- Semantic Neighbors
  - Ideas proposed for related problems in the training set
- KG Neighbors
  - Neighbors of the seed term in background KG
- Citation Neighbors
  - Cited paper title of given background context

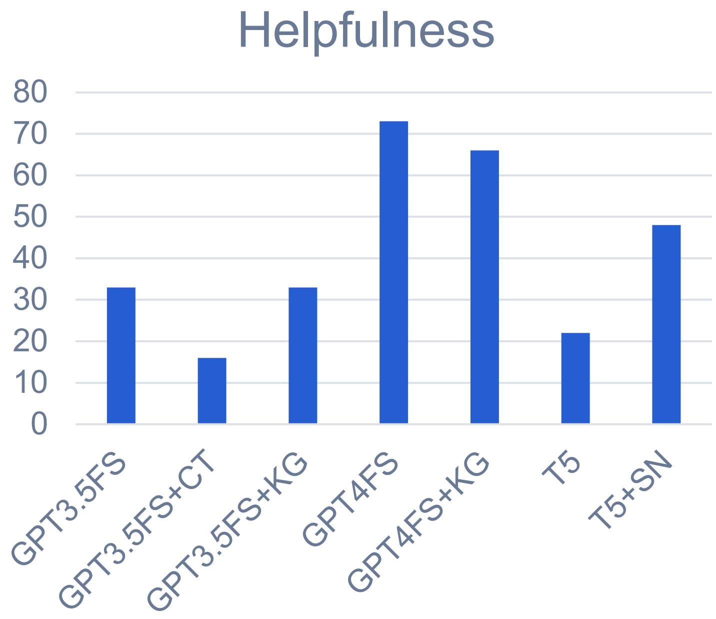
Iterative Novelty Boosting



- We boost novelty iteratively by
  1. retrieving related work from literature reference examples
  2. measuring degree of novelty
  3. instructing the model to update idea to be more novel with respect to reference examples, conditioning on background context

Human Evaluation

- Comparing Outputs across Model Variants
  - Evaluate the generated hypothesis by considering each output's relevance to the context, novelty, clarity, and whether the idea is reasonable
  - GPT4FS and GPT4FS+KG **outperform** other models by a wide margin
  - GPT4 outputs tended to be **longer**, which may partially explain higher human preference
- Comparisons to Real Papers
  - The results are ranked according to the level of technical detail and innovation in comparison to each other and ground truth
  - 48% GPT4FS+KG shows **higher** technical detail
  - 45% GPT4FS+KG is **more novel**
  - Original ground truth ideas have **significantly higher** technical level and novelty in 85% of comparisons



Human Evaluation: Iterative Novelty Boosting

- Compare the novelty-enhanced results against the previous generated ideas
- Examine the new terms added after filtering stopwords and generic words
- Ideas after novelty iterations are **longer** than initial ideas
- For ideas not considered more novel after applying our method, we do not observe a drop in novelty: the method either increases or maintains novelty

Type	**GPT4FS**	**+SN**	**+CT**	**+KG**
1st Novelty Δ (%)	+54.4	**+55.6**	+47.8	+46.7
2nd Novelty Δ (%)	-	+57.8	-	-
1st new terms Δ	**+23.1**	+22.8	+22.1	+21.9
2nd new terms Δ	-	+21.5	-	-

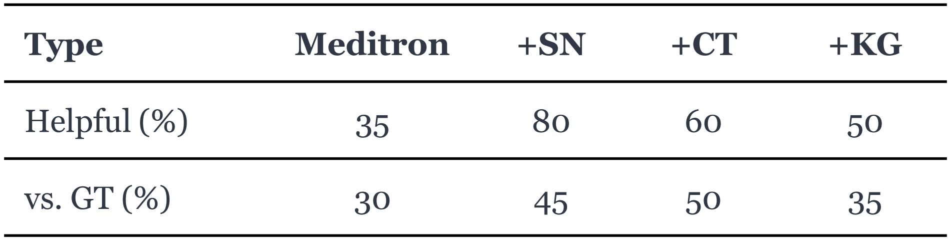
Qualitative Analysis

Input	*seed term*: **speech unit boundaries ;** *context* (abridged): … generate partial sentence translation given a streaming speech input. existing approaches … break the acoustic units in speech, as boundaries between acoustic units in speech are not even…
Initial Idea	A pause prediction model to identify **speech unit boundaries** …
Iteration 1	A method that leverages acoustic and linguistic features to predict **speech unit boundaries** dynamically, ensuring smooth transitions … differs from the existing research as it combines both acoustic properties and linguistic context … adapting to variations in speaker characteristics, speaking styles, and languages.
Iteration 2	A novel method called Adaptive **Speech Unit Boundary** Detection (ASUBD) … a combination of attention mechanisms to focus on relevant acoustic and linguistic features and reinforcement learning to guide the system to make optimal predictions of unit boundaries based on previous decisions…
Ground Truth	… an efficient monotonic segmentation module … accumulate acoustic information incrementally and detect proper **speech unit boundaries**.

Dong, Q., Zhu, Y., Wang, M., & Li, L. (2021). Learning when to translate for streaming speech. ACL 2022.

Case Study: Biomedical

- Collect a dataset from PubMed papers from 1988 to 2024 and construct a biomedical dataset, including 4,767 papers before 2023/02, 642 papers from 2023/02 to 2023/08, and 299 papers after 2023/08



Type	Meditron	+SN	+CT	+KG
Helpful (%)	35	80	60	50
vs. GT (%)	30	45	50	35

Chen, Z., Cano, A. H., Romanou, A., Bonnet, A., Matoba, K., Salvi, F., ... & Bosselut, A. (2023). Meditron-70b: Scaling medical pretraining for large language models. arXiv preprint arXiv:2311.16079.

Case Study: Biomedical

Input	*seed term*: **ARO10;** *context* (abridged): Huangjiu is known for its unique aroma, primarily attributed to its high concentration of b-phenylethanol (ranging from 40 to 130 mg/L).
Meditron	We then selected six key genes involved in the phenylalanine metabolism pathway and predicted that their product flux may affect the synthesis of b-phenylethanol.
+ CT	We found that the key amino acid residue that controls the activity of **Aro10p** was not conserved in wine yeast strains, which may explain the lower b-phenylethanol production in wine fermentation compared with that in Chinese huangjiu.
+ KG	Both target genes, SSA1 and **ARO10**, were deleted using the CRISPR-Cas9 genome editing system.
+ SN	Herein, we report that the key barrier for b-phenylethanol production in Huangjiu is **ARO10**, the only bi-functional amino acid decarboxylase in Saccharomyces cerevisiae.

Chen, Z., Cano, A. H., Romanou, A., Bonnet, A., Matoba, K., Salvi, F., ... & Bosselut, A. (2023). Meditron-70b: Scaling medical pretraining for large language models. arXiv preprint arXiv:2311.16079.

Conclusion

- Propose the first contextualized literature discovery task in which model input background contexts (e.g., problems, experimental settings, goals), and output *natural language ideas* grounded in literature
- Propose a novel approach that uses retrieval of *inspirations* from past scientific papers, and explicitly optimizes for novelty by iteratively comparing to prior papers and updating idea suggestions until sufficient novelty is achieved
- Design extensive evaluation experiments using human annotators with domain expertise to assess relevance, utility, novelty, and technical depth

Code and Data are public at: https://github.com/EagleW/Scientific-Inspiration-Machines-Optimized-for-Novelty

Thank you!

Code and Data are public at:
https://github.com/EagleW/Scientific-Inspiration-Machines-Optimized-for-Novelty