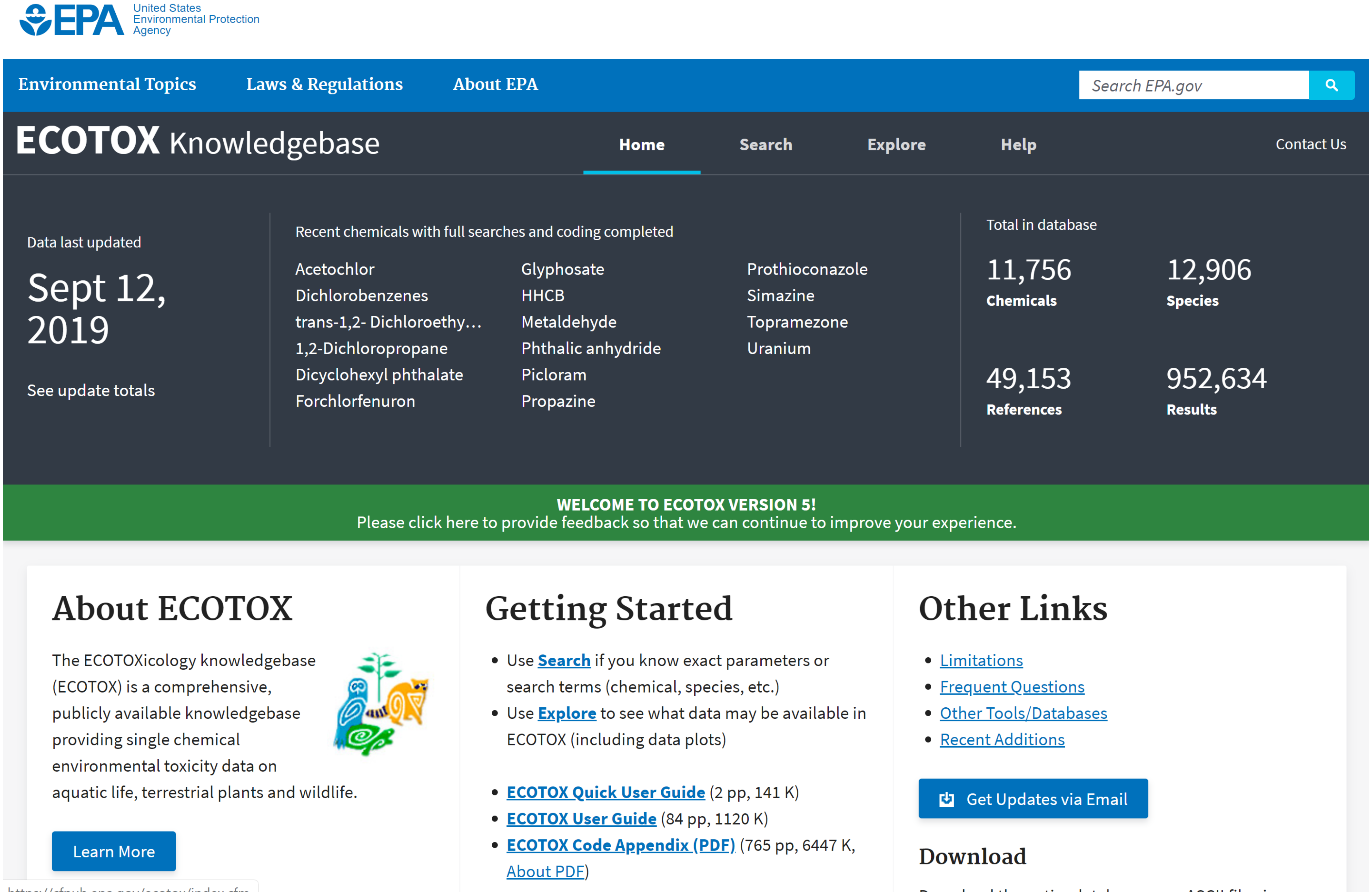
Using Deep Learning and Active Learning Methods to Streamline Literature Curation for the ECOTOXicology knowledgebase

Brian E Howard (Sciome)
Ruchir Shah (Sciome)
Jennifer Olker (EPA)
Colleen Elonen (EPA)
Dale Hoff (EPA)



EPA United States Environmental Protection Agency

ECOTOX Knowledgebase

Data last updated

Sept 12, 2019

See update totals

Recent chemicals with full searches and coding completed

Acetochlor
Dichlorobenzenes
trans-1,2- Dichloroethy…
1,2-Dichloropropane
Dicyclohexyl phthalate
Forchlorfenuron
Glyphosate
HHCB
Metaldehyde
Phthalic anhydride
Picloram
Propazine
Prothioconazole
Simazine
Topramezone
Uranium

Total in database

11,756 **Chemicals**	12,906 **Species**
49,153 **References**	952,634 **Results**

WELCOME TO ECOTOX VERSION 5!
Please click here to provide feedback so that we can continue to improve your experience.

About ECOTOX

The ECOTOXicology knowledgebase (ECOTOX) is a comprehensive, publicly available knowledgebase providing single chemical environmental toxicity data on aquatic life, terrestrial plants and wildlife.

Getting Started

- Use **Search** if you know exact parameters or search terms (chemical, species, etc.)
- Use **Explore** to see what data may be available in ECOTOX (including data plots)

- **ECOTOX Quick User Guide** (2 pp, 141 K)
- **ECOTOX User Guide** (84 pp, 1120 K)
- **ECOTOX Code Appendix (PDF)** (765 pp, 6447 K, About PDF)

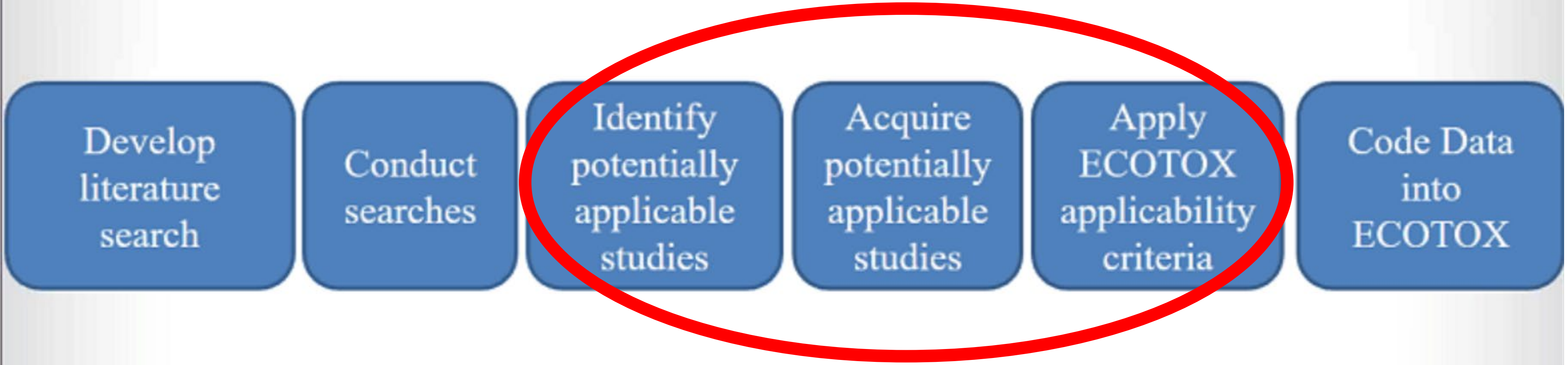
Other Links

- Limitations
- Frequent Questions
- Other Tools/Databases
- Recent Additions

Download

https://cfpub.epa.gov/ecotox/index.cfm

ECOTOX Pipeline



Develop literature search

Conduct searches

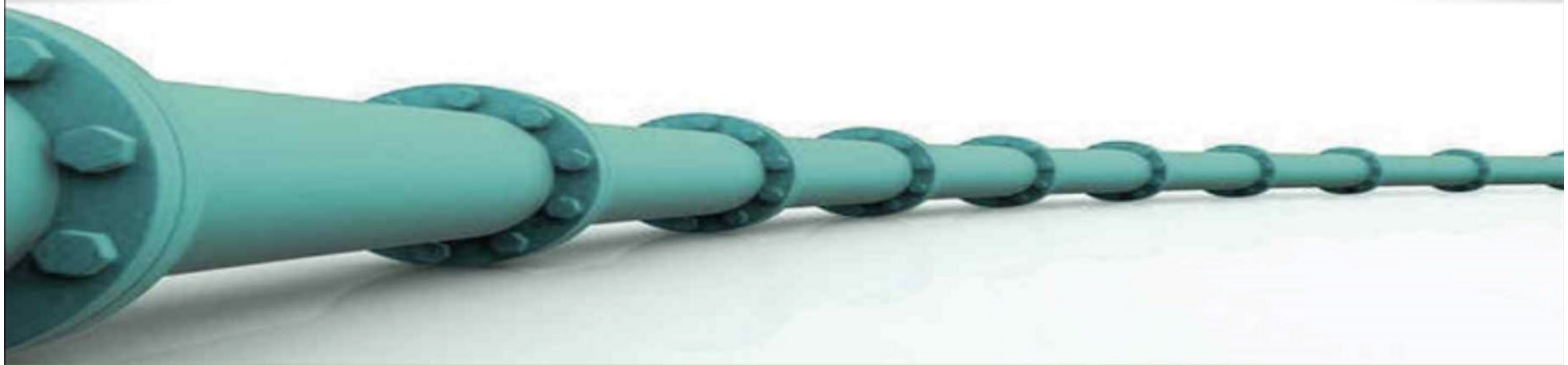
Identify potentially applicable studies

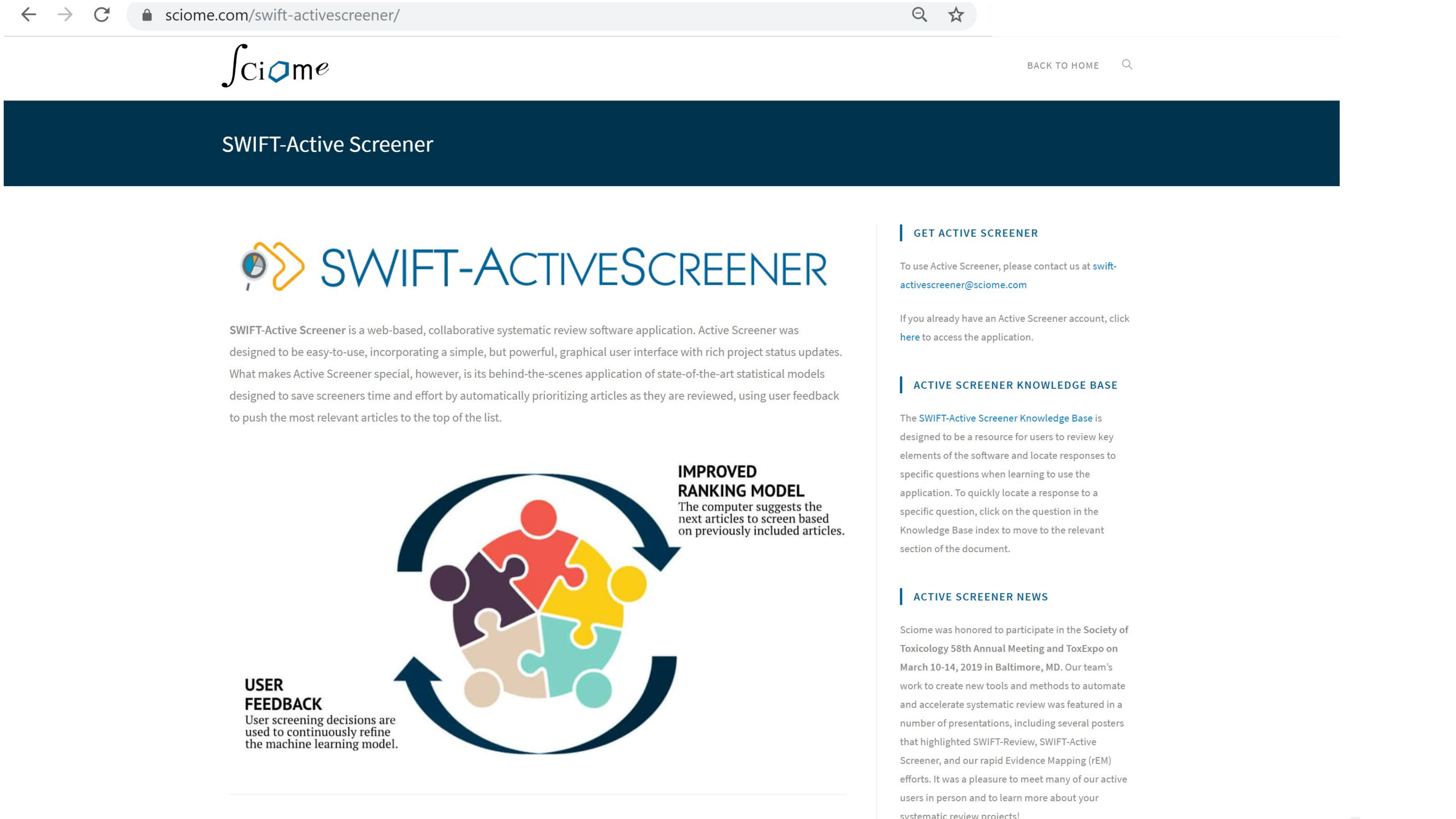
Acquire potentially applicable studies

Apply ECOTOX applicability criteria

Code Data into ECOTOX

Systematic Review and Data Curation





SWIFT-Active Screener

SWIFT-ACTIVESCREENER

SWIFT-Active Screener is a web-based, collaborative systematic review software application. Active Screener was designed to be easy-to-use, incorporating a simple, but powerful, graphical user interface with rich project status updates. What makes Active Screener special, however, is its behind-the-scenes application of state-of-the-art statistical models designed to save screeners time and effort by automatically prioritizing articles as they are reviewed, using user feedback to push the most relevant articles to the top of the list.

IMPROVED RANKING MODEL
The computer suggests the next articles to screen based on previously included articles.

USER FEEDBACK
User screening decisions are used to continuously refine the machine learning model.

GET ACTIVE SCREENER

To use Active Screener, please contact us at swift-activescreener@sciome.com

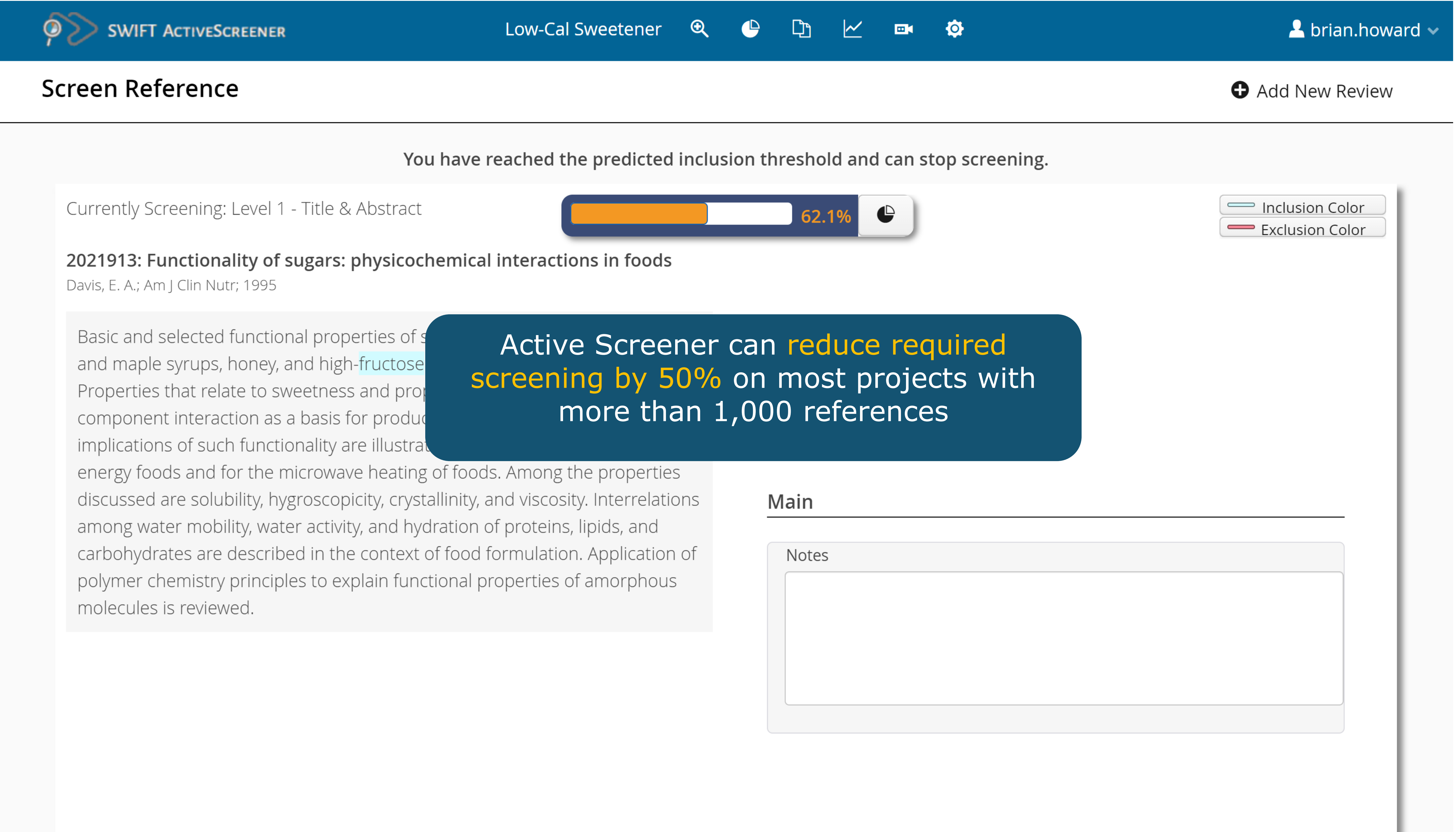
If you already have an Active Screener account, click here to access the application.

ACTIVE SCREENER KNOWLEDGE BASE

The SWIFT-Active Screener Knowledge Base is designed to be a resource for users to review key elements of the software and locate responses to specific questions when learning to use the application. To quickly locate a response to a specific question, click on the question in the Knowledge Base index to move to the relevant section of the document.

ACTIVE SCREENER NEWS

Sciome was honored to participate in the **Society of Toxicology 58th Annual Meeting and ToxExpo on March 10-14, 2019 in Baltimore, MD.** Our team's work to create new tools and methods to automate and accelerate systematic review was featured in a number of presentations, including several posters that highlighted SWIFT-Review, SWIFT-Active Screener, and our rapid Evidence Mapping (rEM) efforts. It was a pleasure to meet many of our active users in person and to learn more about your systematic review projects!



Screen Reference

You have reached the predicted inclusion threshold and can stop screening.

Currently Screening: Level 1 - Title & Abstract

2021913: Functionality of sugars: physicochemical interactions in foods

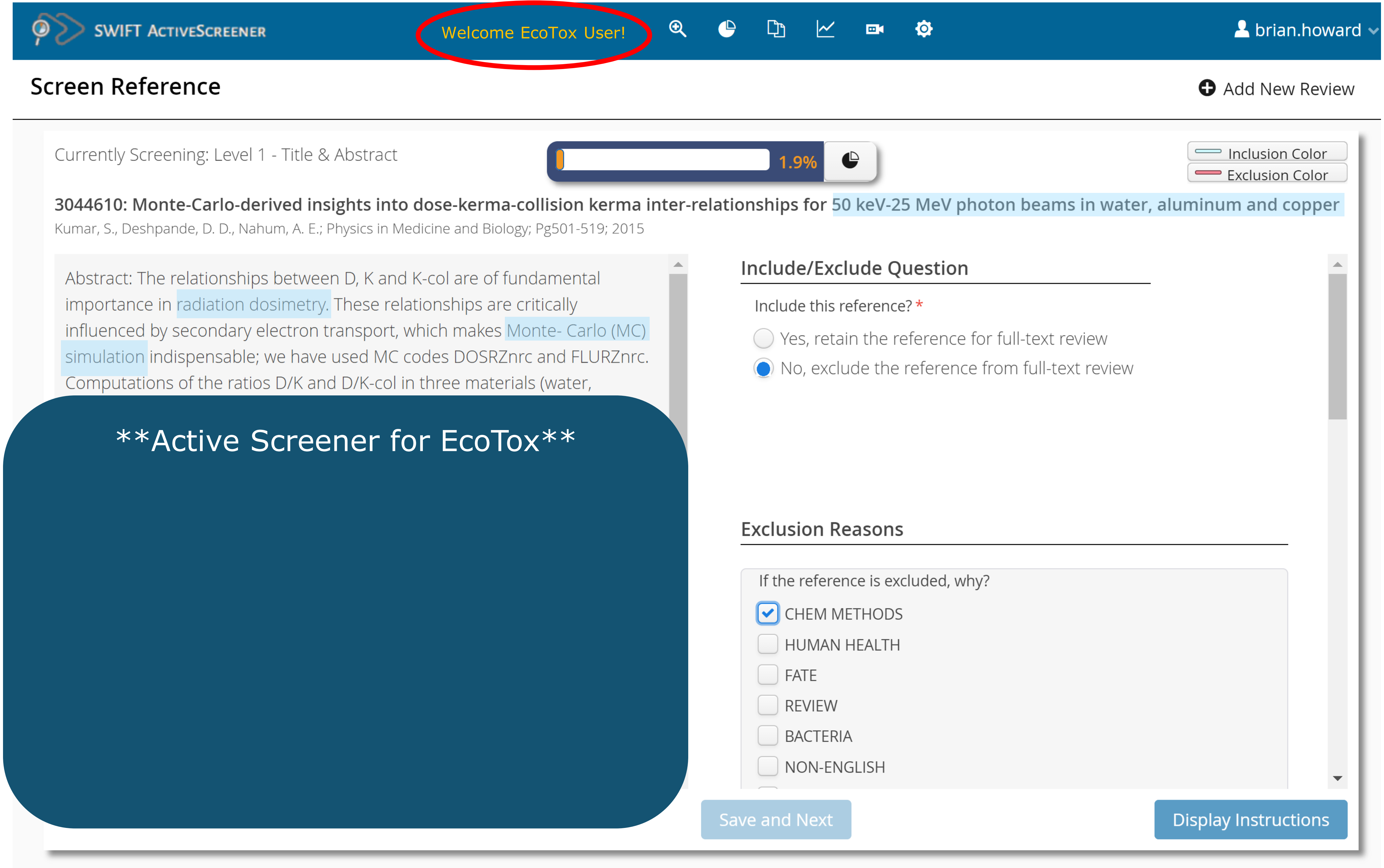
Davis, E. A.; Am J Clin Nutr; 1995

Basic and selected functional properties of s
and maple syrups, honey, and high-fructose
Properties that relate to sweetness and prop
component interaction as a basis for produc
implications of such functionality are illustra
energy foods and for the microwave heating of foods. Among the properties discussed are solubility, hygroscopicity, crystallinity, and viscosity. Interrelations among water mobility, water activity, and hydration of proteins, lipids, and carbohydrates are described in the context of food formulation. Application of polymer chemistry principles to explain functional properties of amorphous molecules is reviewed.

Active Screener can reduce required screening by 50% on most projects with more than 1,000 references

Main

Notes



Screen Reference

Currently Screening: Level 1 - Title & Abstract

3044610: Monte-Carlo-derived insights into dose-kerma-collision kerma inter-relationships for 50 keV-25 MeV photon beams in water, aluminum and copper

Kumar, S., Deshpande, D. D., Nahum, A. E.; Physics in Medicine and Biology; Pg501-519; 2015

Abstract: The relationships between D, K and K-col are of fundamental importance in radiation dosimetry. These relationships are critically influenced by secondary electron transport, which makes Monte- Carlo (MC) simulation indispensable; we have used MC codes DOSRZnrc and FLURZnrc. Computations of the ratios D/K and D/K-col in three materials (water,

Active Screener for EcoTox

Include/Exclude Question

Include this reference? *

○ Yes, retain the reference for full-text review

● No, exclude the reference from full-text review

Exclusion Reasons

If the reference is excluded, why?

☑ CHEM METHODS

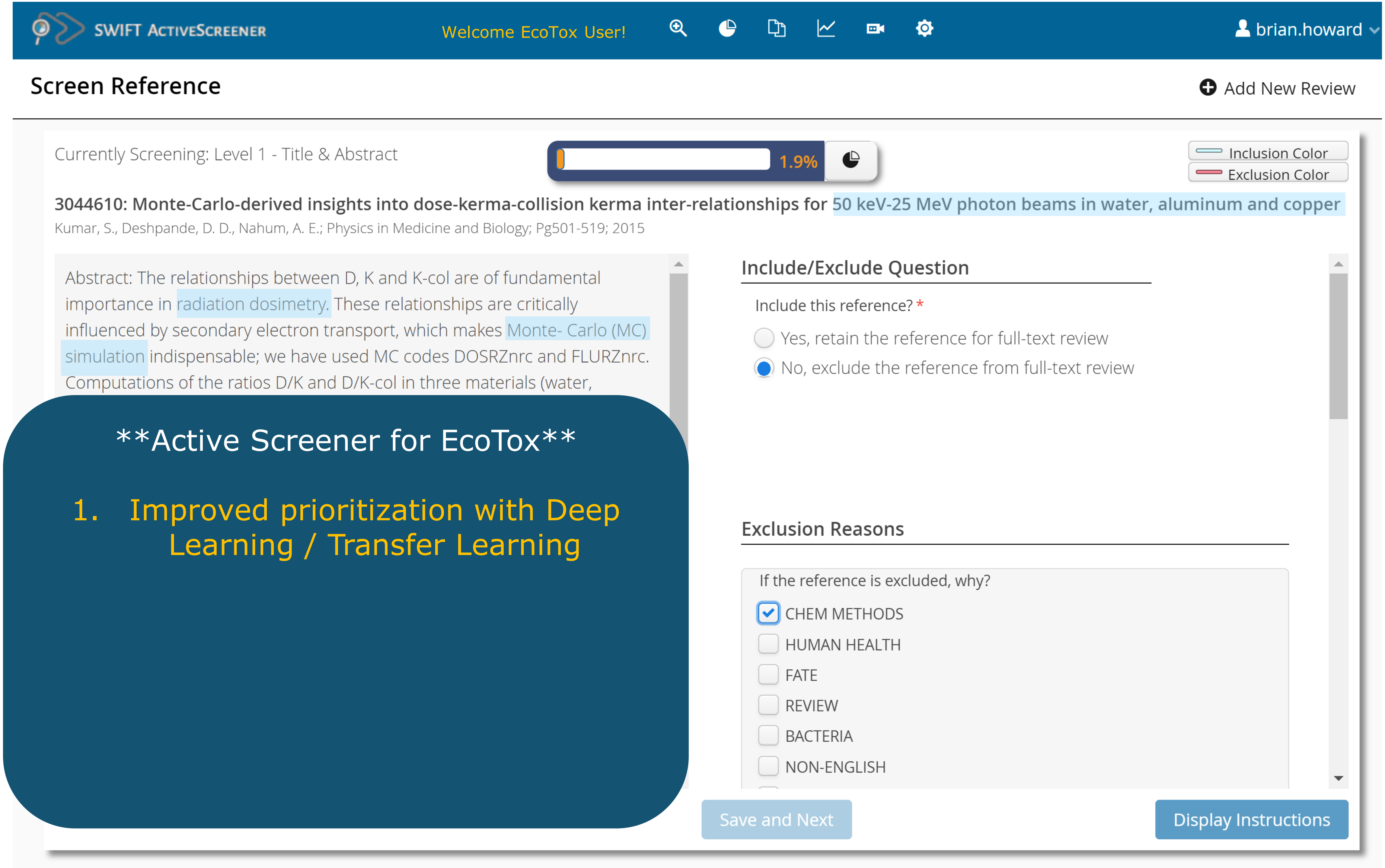
□ HUMAN HEALTH

□ FATE

□ REVIEW

□ BACTERIA

□ NON-ENGLISH



Screen Reference

Currently Screening: Level 1 - Title & Abstract

3044610: Monte-Carlo-derived insights into dose-kerma-collision kerma inter-relationships for 50 keV-25 MeV photon beams in water, aluminum and copper

Kumar, S., Deshpande, D. D., Nahum, A. E.; Physics in Medicine and Biology; Pg501-519; 2015

Abstract: The relationships between D, K and K-col are of fundamental importance in radiation dosimetry. These relationships are critically influenced by secondary electron transport, which makes Monte- Carlo (MC) simulation indispensable; we have used MC codes DOSRZnrc and FLURZnrc. Computations of the ratios D/K and D/K-col in three materials (water,

Include/Exclude Question

Include this reference? *

○ Yes, retain the reference for full-text review

● No, exclude the reference from full-text review

Exclusion Reasons

If the reference is excluded, why?

☑ CHEM METHODS

□ HUMAN HEALTH

□ FATE

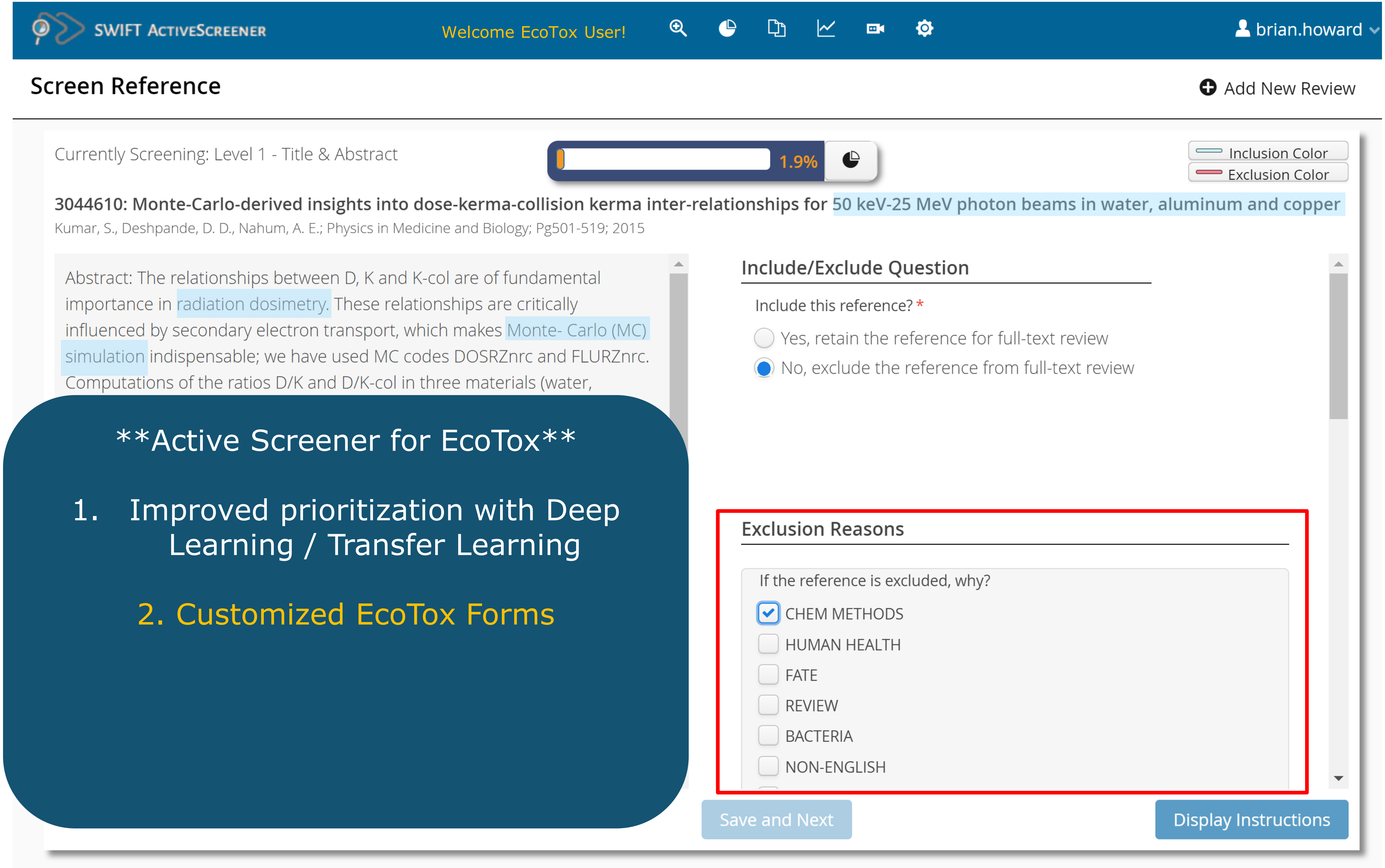
□ REVIEW

□ BACTERIA

□ NON-ENGLISH

Active Screener for EcoTox

1. Improved prioritization with Deep Learning / Transfer Learning



Screen Reference

Currently Screening: Level 1 - Title & Abstract

3044610: Monte-Carlo-derived insights into dose-kerma-collision kerma inter-relationships for 50 keV-25 MeV photon beams in water, aluminum and copper

Kumar, S., Deshpande, D. D., Nahum, A. E.; Physics in Medicine and Biology; Pg501-519; 2015

Abstract: The relationships between D, K and K-col are of fundamental importance in radiation dosimetry. These relationships are critically influenced by secondary electron transport, which makes Monte- Carlo (MC) simulation indispensable; we have used MC codes DOSRZnrc and FLURZnrc. Computations of the ratios D/K and D/K-col in three materials (water,

Include/Exclude Question

Include this reference? *

○ Yes, retain the reference for full-text review

● No, exclude the reference from full-text review

Exclusion Reasons

If the reference is excluded, why?

☑ CHEM METHODS

□ HUMAN HEALTH

□ FATE

□ REVIEW

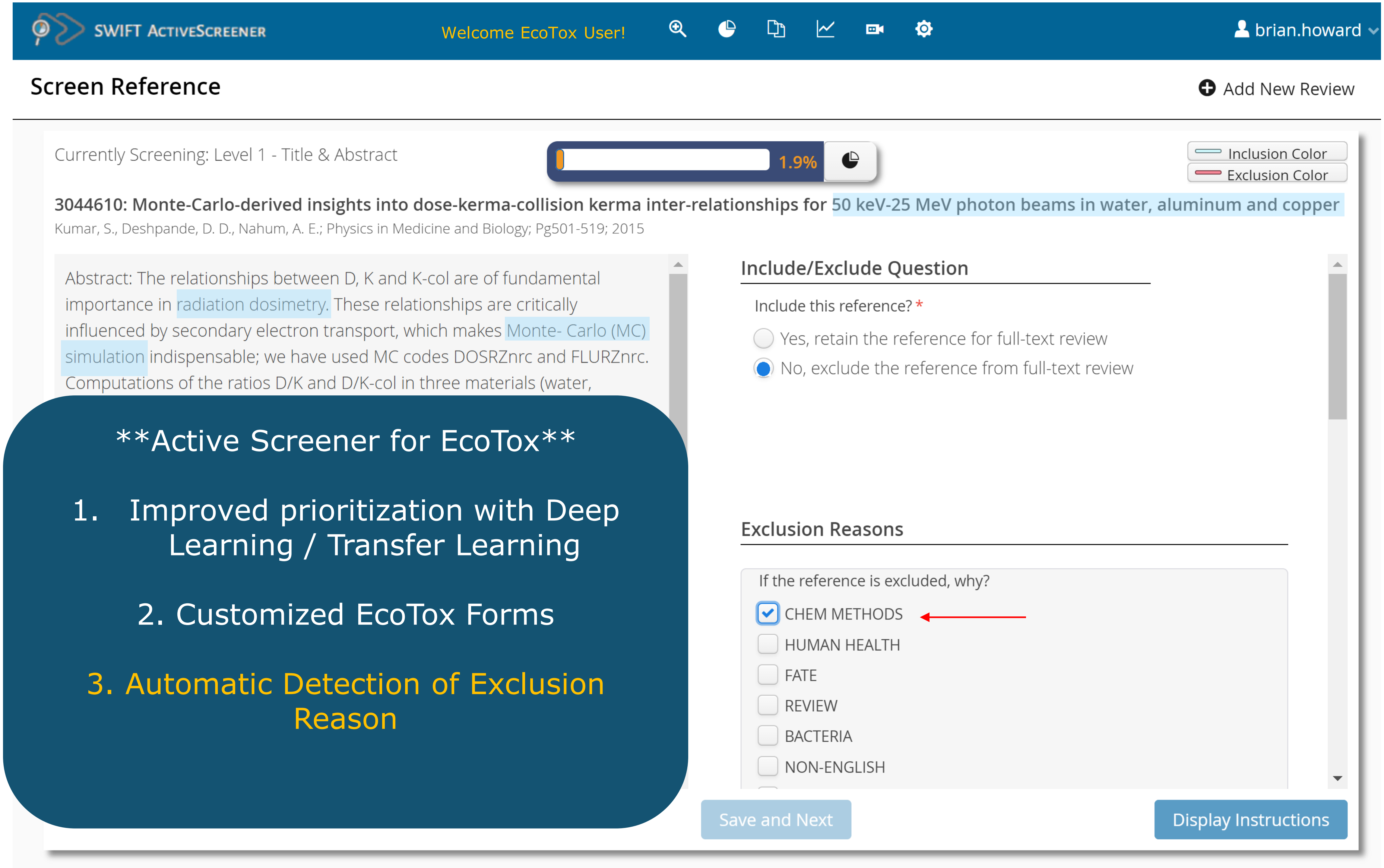
□ BACTERIA

□ NON-ENGLISH

Active Screener for EcoTox

1. Improved prioritization with Deep Learning / Transfer Learning

2. Customized EcoTox Forms



Screen Reference

Currently Screening: Level 1 - Title & Abstract

3044610: Monte-Carlo-derived insights into dose-kerma-collision kerma inter-relationships for 50 keV-25 MeV photon beams in water, aluminum and copper

Kumar, S., Deshpande, D. D., Nahum, A. E.; Physics in Medicine and Biology; Pg501-519; 2015

Abstract: The relationships between D, K and K-col are of fundamental importance in radiation dosimetry. These relationships are critically influenced by secondary electron transport, which makes Monte- Carlo (MC) simulation indispensable; we have used MC codes DOSRZnrc and FLURZnrc. Computations of the ratios D/K and D/K-col in three materials (water,

Include/Exclude Question

Include this reference? *

○ Yes, retain the reference for full-text review

● No, exclude the reference from full-text review

Exclusion Reasons

If the reference is excluded, why?

☑ CHEM METHODS

□ HUMAN HEALTH

□ FATE

□ REVIEW

□ BACTERIA

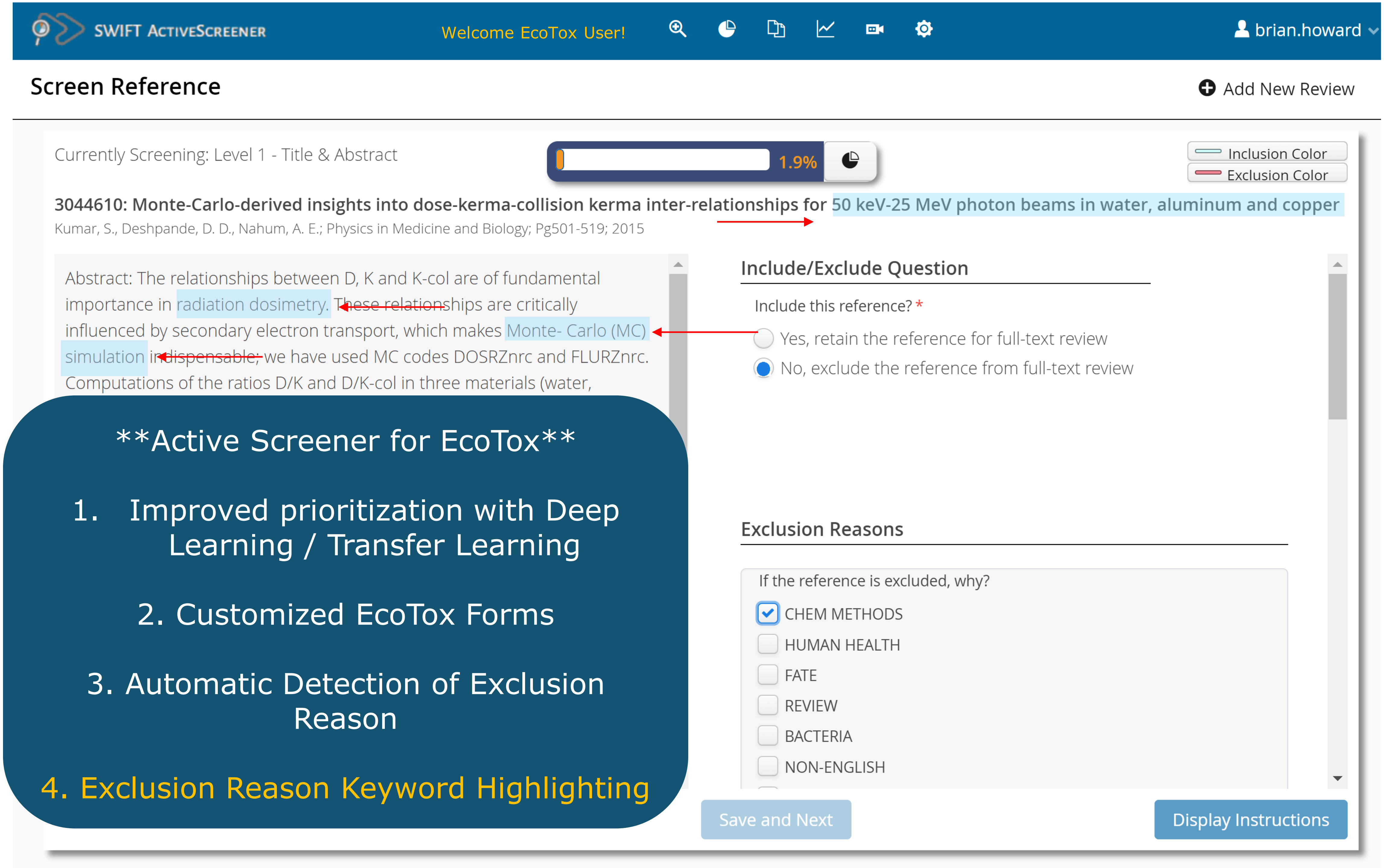
□ NON-ENGLISH

Active Screener for EcoTox

1. Improved prioritization with Deep Learning / Transfer Learning

2. Customized EcoTox Forms

3. Automatic Detection of Exclusion Reason



Screen Reference

Currently Screening: Level 1 - Title & Abstract

3044610: Monte-Carlo-derived insights into dose-kerma-collision kerma inter-relationships for 50 keV-25 MeV photon beams in water, aluminum and copper

Kumar, S., Deshpande, D. D., Nahum, A. E.; Physics in Medicine and Biology; Pg501-519; 2015

Abstract: The relationships between D, K and K-col are of fundamental importance in radiation dosimetry. ~~These relationships~~ are critically influenced by secondary electron transport, which makes Monte- Carlo (MC) simulation in~~dispensable;~~ we have used MC codes DOSRZnrc and FLURZnrc. Computations of the ratios D/K and D/K-col in three materials (water,

Include/Exclude Question

Include this reference? *

○ Yes, retain the reference for full-text review

● No, exclude the reference from full-text review

Exclusion Reasons

If the reference is excluded, why?

☑ CHEM METHODS

□ HUMAN HEALTH

□ FATE

□ REVIEW

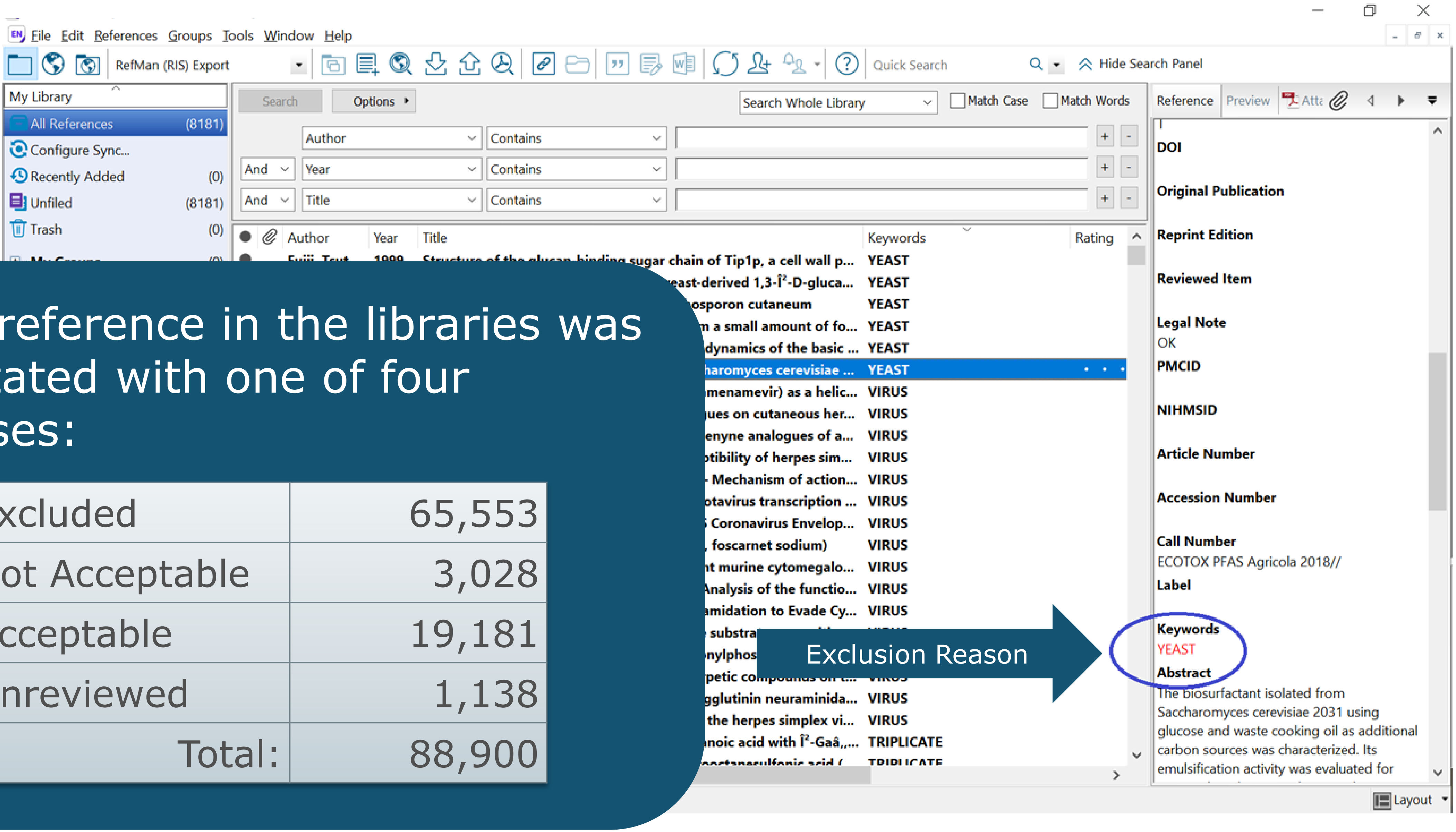
□ BACTERIA

□ NON-ENGLISH

Active Screener for EcoTox

1. Improved prioritization with Deep Learning / Transfer Learning
2. Customized EcoTox Forms
3. Automatic Detection of Exclusion Reason
4. Exclusion Reason Keyword Highlighting

Existing Datasets



My Library

All References	(8181)
Configure Sync...	
Recently Added	(0)
Unfiled	(8181)
Trash	(0)
My Groups	(0)

Search Whole Library ☐ Match Case ☐ Match Words

	Author	Contains	
And	Year	Contains	
And	Title	Contains	

Author	Year	Title	Keywords	Rating
Fujii, Tsut	1999	Structure of the glucan-binding sugar chain of Tip1p, a cell wall p...	YEAST	
		east-derived 1,3-Î²-D-gluca...	YEAST	
		osporon cutaneum	YEAST	
		m a small amount of fo...	YEAST	
		dynamics of the basic ...	YEAST	
		haromyces cerevisiae ...	YEAST	
		menamevir) as a helic...	VIRUS	
		ues on cutaneous her...	VIRUS	
		enyne analogues of a...	VIRUS	
		otibility of herpes sim...	VIRUS	
		- Mechanism of action...	VIRUS	
		otavirus transcription ...	VIRUS	
		Coronavirus Envelop...	VIRUS	
		, foscarnet sodium)	VIRUS	
		nt murine cytomegalo...	VIRUS	
		Analysis of the functio...	VIRUS	
		amidation to Evade Cy...	VIRUS	
		substra		
		nylphos		
		petic co		
		gglutinin neuraminida...	VIRUS	
		the herpes simplex vi...	VIRUS	
		noic acid with Î²-Gaâ,...	TRIPLICATE	
		octanesulfonic acid (	TRIPLICATE	

DOI

Original Publication

Reprint Edition

Reviewed Item

Legal Note
OK

PMCID

NIHMSID

Article Number

Accession Number

Call Number
ECOTOX PFAS Agricola 2018//

Label

Keywords
YEAST

Abstract
The biosurfactant isolated from Saccharomyces cerevisiae 2031 using glucose and waste cooking oil as additional carbon sources was characterized. Its emulsification activity was evaluated for

Exclusion Reason

Each reference in the libraries was annotated with one of four statuses:

Excluded	65,553
Not Acceptable	3,028
Acceptable	19,181
Unreviewed	1,138
Total:	88,900

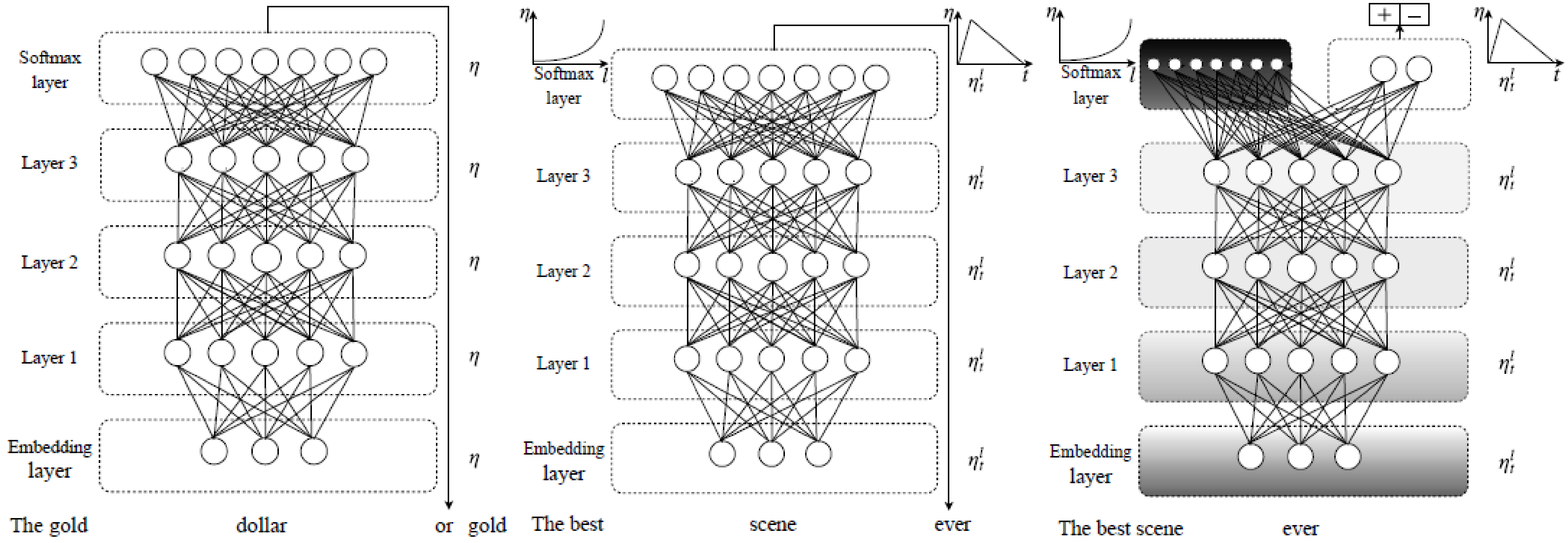
Existing Datasets

- Excluded articles also were associated with a reason for exclusion.

- The top 20 reasons make up over 95% of the data. The remaining terms were combined as an "Other" category.

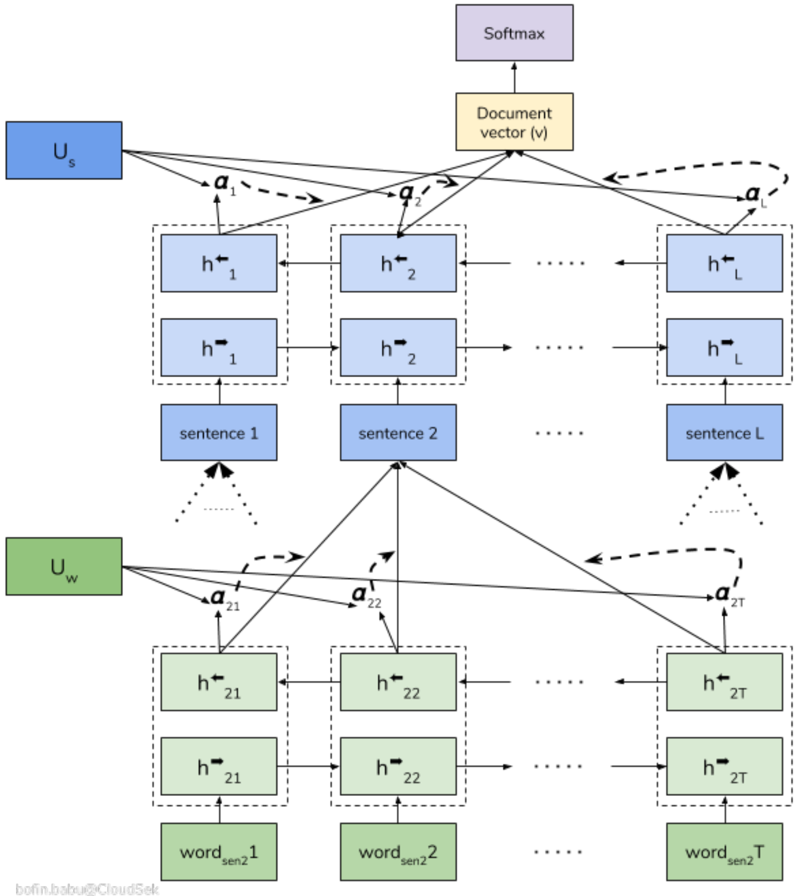
Exclusion Reason	Refs	Percentage
HUMAN HEALTH	19609	30.41%
CHEM METHODS	16745	25.97%
NO TOXICANT	8074	12.52%
FATE	5184	8.04%
BACTERIA	2961	4.59%
REVIEW	2251	3.49%
SURVEY	1696	2.63%
MIXTURE	1101	1.71%
NON-ENGLISH	1003	1.56%
ABSTRACT	939	1.46%
IN VITRO	805	1.25%
OTHER	701	1.09%
……..	……..	……..
BIOLOGICAL TOXICANT	105	0.16%
	64,480	

ULMFit Classifier (Howard and Ruder, 2018)



(a) LM pre-training (b) LM fine-tuning (c) Classifier fine-tuning

Adding Attention to ULMFit



Sentence attention

U_s: sentence context vector

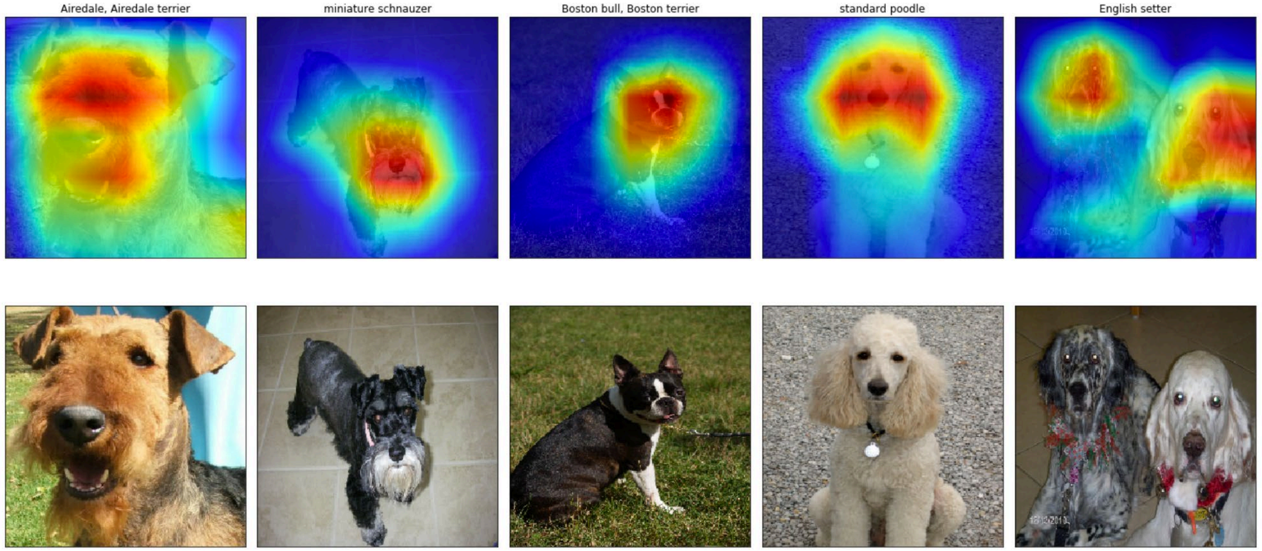
Sentence encoder

Word attention

α : importance weight
U_w : word context vector (*what is the informative word?*)

Word encoder

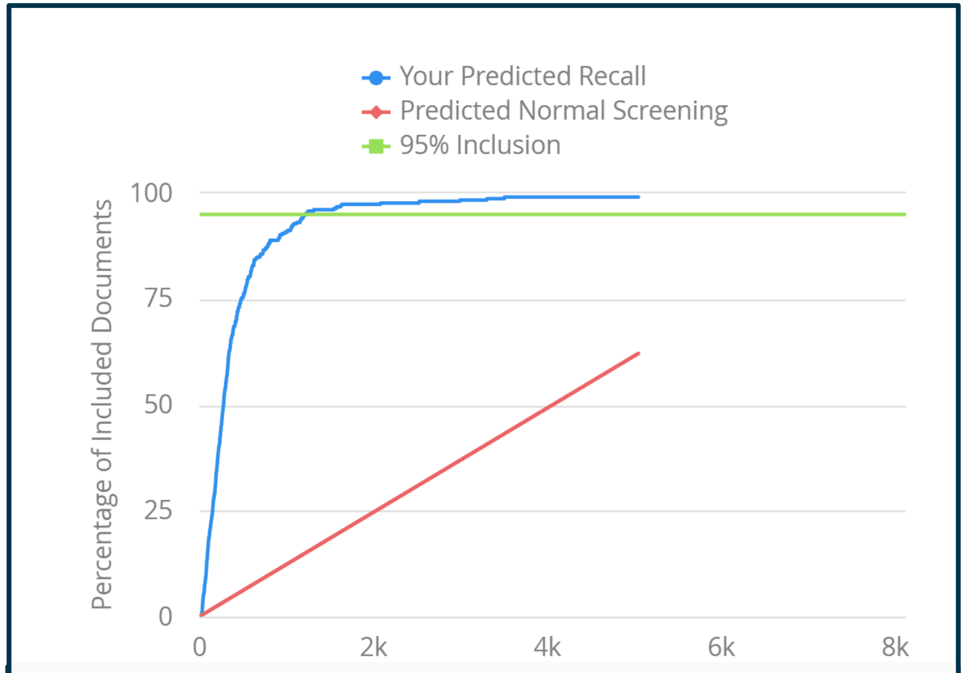
$\overrightarrow{h}$: forward hidden state
$\overleftarrow{h}$: backward hidden state
$\text{word}_{\text{sen2}}1$: The first word in sentence 2

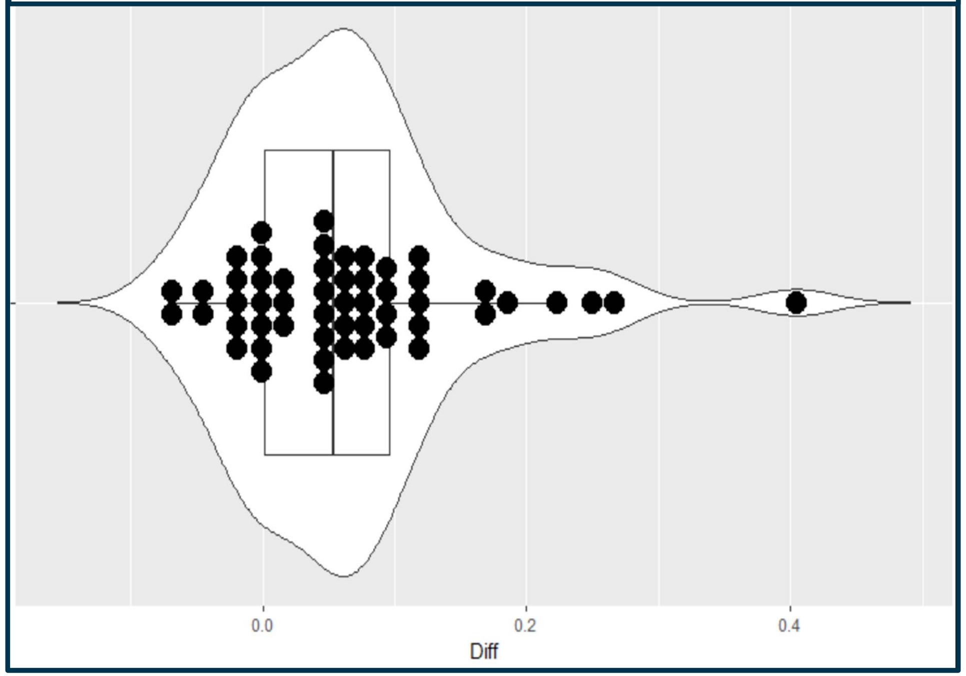


Results: Acceptable / Not Acceptable

Evaluated whether machine learning can be used to classify documents as Acceptable vs Not Acceptable / Excluded and found that:

- Using Active Screener can save users 50% of screening effort for many datasets.
- Augmenting standard model with pretrained model via transfer learning provides additional benefits (mean improvement of 6.5% WSS over the standard Active Screener prioritization model, but several datasets had significantly larger gains).



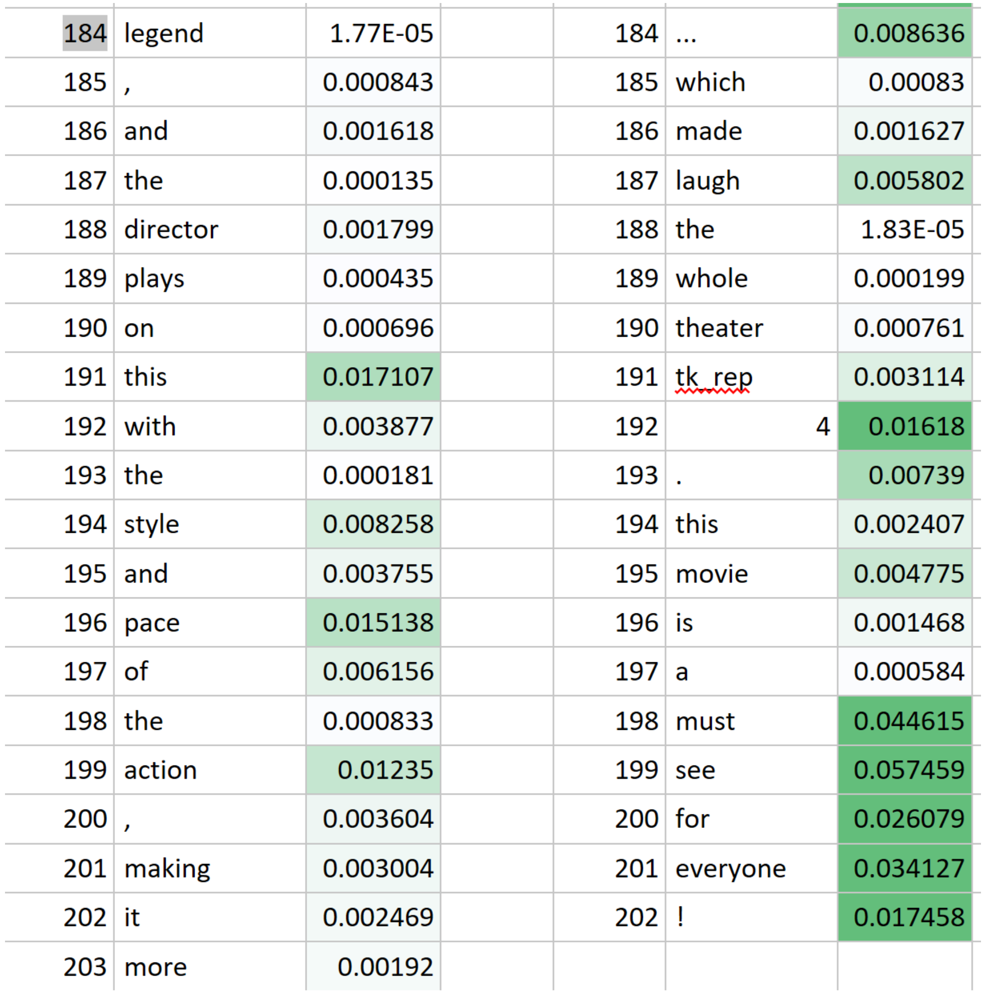


Results: Exclusion Reason

Label	Total Refs	% Refs	Accuracy	Recall	Precision	F1
CHEM METHODS	7022	27.53%	89.77%	74.09%	77.47%	75.74%
HUMAN HEALTH	4752	18.63%	86.43%	69.70%	61.95%	65.60%
FATE	2875	11.27%	95.56%	69.86%	61.39%	65.35%
REVIEW	1800	7.06%	94.23%	60.78%	62.85%	61.80%
BACTERIA	1359	5.33%	95.58%	54.64%	35.69%	43.18%
NON-ENGLISH	940	3.68%	97.53%	68.33%	74.38%	71.23%
SURVEY	914	3.58%	95.89%	49.80%	59.15%	54.08%
MIXTURE	809	3.17%	97.12%	61.54%	55.03%	58.10%
IN VITRO	805	3.16%	95.73%	47.37%	54.55%	50.70%
ABSTRACT	791	3.10%	98.06%	61.19%	54.67%	57.75%
REFS CHECKED	697	2.73%	97.51%	68.84%	57.93%	62.91%
NO SOURCE	370	1.45%	96.25%	57.84%	64.46%	60.97%
NO CONC	336	1.32%	97.89%	51.49%	46.43%	48.83%
MODELING	284	1.11%	98.78%	63.64%	30.43%	41.18%
NO EFFECT	253	0.99%	97.93%	6.86%	17.95%	9.93%
METHODS	249	0.98%	99.47%	60.00%	68.57%	64.00%
FOOD	220	0.86%	99.03%	35.71%	38.46%	37.04%
PUBL AS	157	0.62%	99.25%	37.84%	41.18%	39.44%
NO DURATION	142	0.56%	99.49%	50.00%	60.00%	54.55%
YEAST	111	0.44%	99.55%	70.59%	83.72%	76.60%
OTHER	625	2.44%	95.76%	10.73%	27.85%	15.49%

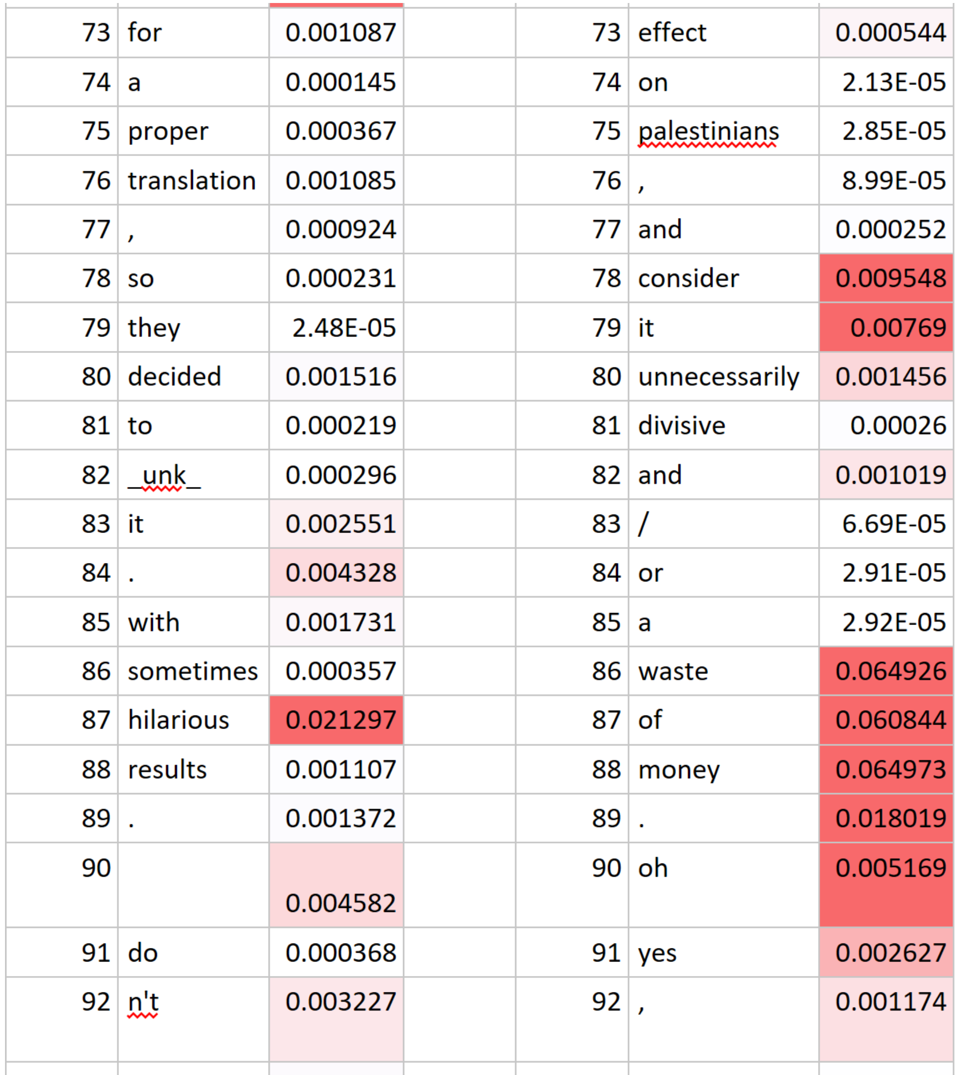
Adding Attention (IMDB Example)

Positive Review



184	legend	1.77E-05	184	...	0.008636
185	,	0.000843	185	which	0.00083
186	and	0.001618	186	made	0.001627
187	the	0.000135	187	laugh	0.005802
188	director	0.001799	188	the	1.83E-05
189	plays	0.000435	189	whole	0.000199
190	on	0.000696	190	theater	0.000761
191	this	0.017107	191	tk_rep	0.003114
192	with	0.003877	192	4	0.01618
193	the	0.000181	193	.	0.00739
194	style	0.008258	194	this	0.002407
195	and	0.003755	195	movie	0.004775
196	pace	0.015138	196	is	0.001468
197	of	0.006156	197	a	0.000584
198	the	0.000833	198	must	0.044615
199	action	0.01235	199	see	0.057459
200	,	0.003604	200	for	0.026079
201	making	0.003004	201	everyone	0.034127
202	it	0.002469	202	!	0.017458
203	more	0.00192			

Negative Review



73	for	0.001087	73	effect	0.000544
74	a	0.000145	74	on	2.13E-05
75	proper	0.000367	75	palestinians	2.85E-05
76	translation	0.001085	76	,	8.99E-05
77	,	0.000924	77	and	0.000252
78	so	0.000231	78	consider	0.009548
79	they	2.48E-05	79	it	0.00769
80	decided	0.001516	80	unnecessarily	0.001456
81	to	0.000219	81	divisive	0.00026
82	_unk_	0.000296	82	and	0.001019
83	it	0.002551	83	/	6.69E-05
84	.	0.004328	84	or	2.91E-05
85	with	0.001731	85	a	2.92E-05
86	sometimes	0.000357	86	waste	0.064926
87	hilarious	0.021297	87	of	0.060844
88	results	0.001107	88	money	0.064973
89	.	0.001372	89	.	0.018019
90		0.004582	90	oh	0.005169
91	do	0.000368	91	yes	0.002627
92	n't	0.003227	92	,	0.001174

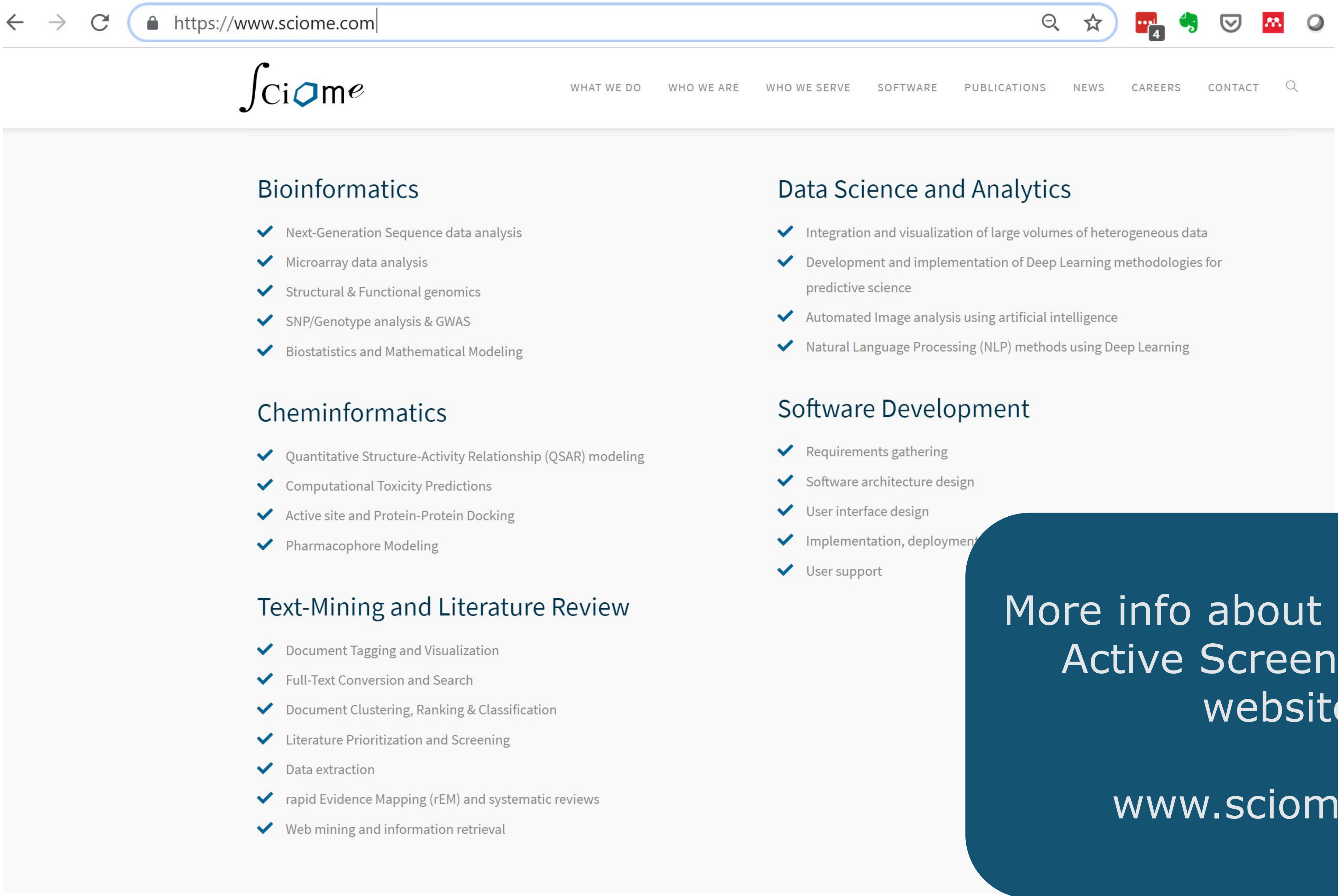
Summary

- Standard Active Screener application saves users 50% screening time
- EcoTox Active Screener uses Deep Learning to:
  - Save an additional 6.5+% screening time
  - Accurately predict exclusion reasons
  - Explain its predictions using attention-highlighting
- Models will continue to improve with more data, and several methodological enhancements are planned

Next Steps

Phase II of the project…

- **Aim 2.1**: **Additional refinements to machine learning models** that can be used to automatically identify, with high precision, those references that can be deemed **non-acceptable/non-applicable** for the EcoTox database, and to **categorize excluded references** according to a selection from a list of pre-defined rationales.
- **Aim 2.2**: Modify **Active Screener** to **operationalize the above models** and to better serve EcoTox data curation pipeline.
- **Aim 2.3**: **Publish** results in a suitable journal or conference.
- **Aim 2.4**: Investigate feasibility of developing models to extract the approximately 4,000 Effects Groups and Measurement Codes from full-text documents.



Bioinformatics

- ✓ Next-Generation Sequence data analysis
- ✓ Microarray data analysis
- ✓ Structural & Functional genomics
- ✓ SNP/Genotype analysis & GWAS
- ✓ Biostatistics and Mathematical Modeling

Cheminformatics

- ✓ Quantitative Structure-Activity Relationship (QSAR) modeling
- ✓ Computational Toxicity Predictions
- ✓ Active site and Protein-Protein Docking
- ✓ Pharmacophore Modeling

Text-Mining and Literature Review

- ✓ Document Tagging and Visualization
- ✓ Full-Text Conversion and Search
- ✓ Document Clustering, Ranking & Classification
- ✓ Literature Prioritization and Screening
- ✓ Data extraction
- ✓ rapid Evidence Mapping (rEM) and systematic reviews
- ✓ Web mining and information retrieval

Data Science and Analytics

- ✓ Integration and visualization of large volumes of heterogeneous data
- ✓ Development and implementation of Deep Learning methodologies for predictive science
- ✓ Automated Image analysis using artificial intelligence
- ✓ Natural Language Processing (NLP) methods using Deep Learning

Software Development

- ✓ Requirements gathering
- ✓ Software architecture design
- ✓ User interface design
- ✓ Implementation, deployment
- ✓ User support

More info about Sciome and Active Screener at our website:

www.sciome.com

ANY
QUESTIONS?