

ON LATENT DOMAIN-SPECIFIC TEXTUAL PREFERENCES IN SOLVING INTERNET-BASED GENERIC TASKS AMONG GRADUATES/YOUNG PROFESSIONALS FROM THREE DOMAINS

ALEXANDER MEHLER¹ MAXIM KONCA¹ MARIE-THERES NAGEL²
ANDY LÜCKING¹ OLGA ZLATKIN-TROITSCHANSKAIA²

¹GOETHE-UNIVERSITÄT FRANKFURT, TEXT TECHNOLOGY LAB

²JOHANNES GUTENBERG-UNIVERSITÄT MAINZ

GEBF, 09–11. 03. 2022



Critical Online Reasoning¹

Critical thinking wrt. online information:

- (i) online information acquisition
- (ii) critical information evaluation
- (iii) reasoning based on evidence, argumentation, and synthesis

- Online study: e-bike and health (German, browser logged)
- Participants: Graduates/young professionals in **medicine, law, and teaching** (domains)

¹D. Molerov et al. (2020). "Assessing University Students' Critical Online Reasoning Ability: A Conceptual and Assessment Framework With Preliminary Evidence". In: Front. Educ. 5, 577843.

Questions

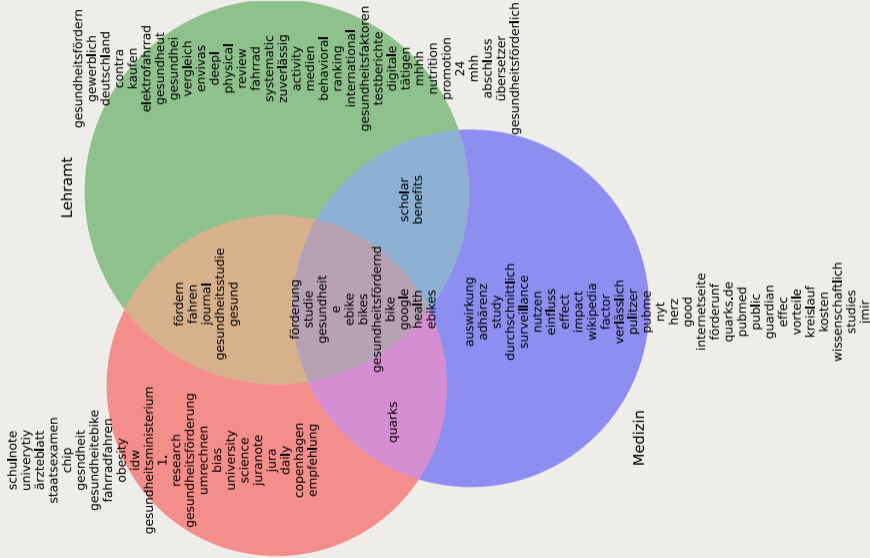
1. domain difference in search behavior?
 2. domain difference among consulted websites?
- @1: search terms and types of consulted websites
 - @2: text-technological classification of websites (text features)

@1: SEARCH VENN DIAGRAM

green = teaching,
red = law,
purple = med.

- mutual search term overlap

➔ Observe the large proportion of **domain-specific search terms**



@1: SEARCH TERM EXAMPLES (SELECTION)

Shared	Only Medicine	Only Teacher	Only Law
<ul style="list-style-type: none">■ förderung (aid)	<ul style="list-style-type: none">■ adhärenz (adherence)	<ul style="list-style-type: none">■ gewerblich (commercial)	<ul style="list-style-type: none">■ ärzteblatt [!]
<ul style="list-style-type: none">■ studie (study)	<ul style="list-style-type: none">■ einfluss (influence)	<ul style="list-style-type: none">■ contra	<ul style="list-style-type: none">■ staatsexamen (state examin.)
<ul style="list-style-type: none">■ gesundheit (health)	<ul style="list-style-type: none">■ wikipedia	<ul style="list-style-type: none">■ kaufen (buy)	<ul style="list-style-type: none">■ obesity (“fatness”)
<ul style="list-style-type: none">■ e-bike (various spellings)	<ul style="list-style-type: none">■ pulitzer	<ul style="list-style-type: none">■ mhh (med. univ.)	<ul style="list-style-type: none">■ gesundheitsministerium (health ministry)
<ul style="list-style-type: none">■ gesundheitsfördernd (health enhancing)	<ul style="list-style-type: none">■ pubmed	<ul style="list-style-type: none">■ nutrition	<ul style="list-style-type: none">■ bias
<ul style="list-style-type: none">■ google	<ul style="list-style-type: none">■ guardian	<ul style="list-style-type: none">■ activity	<ul style="list-style-type: none">■ daily
<ul style="list-style-type: none">■ health	<ul style="list-style-type: none">■ quarks.de	<ul style="list-style-type: none">■ medien (media)	<ul style="list-style-type: none">■ empfehlung (recommend.)
	<ul style="list-style-type: none">■ kosten (cost)	<ul style="list-style-type: none">■ testberichte (test reports)	<ul style="list-style-type: none">■ ...
	<ul style="list-style-type: none">■ jmir	<ul style="list-style-type: none">■ ranking	
	<ul style="list-style-type: none">■ ...	<ul style="list-style-type: none">■ ...	

- shared search terms pretty directly reflect the task
- other ones draw on common ground develop in the domains, but also show individual background knowledge
- different search terms lead to different results and hence visited websites, but which ones?

Mainz (prev. talk): type and reliability

- 2 raters
- kappa = 0.76,
Krippendorff's alpha = 0.759

0.6 < α ≤ 0.8 substantial
agreement

α > 0.8 near-perfect agreement

Wiki-based

1. Wikidata classification
(usually manifold)
2. Wikipedia ("X is a(n) ...")
3. free web search ("X is a(n) ...")

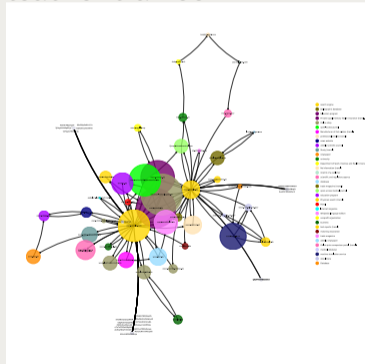
→ our focus today!

examples: search engine, online shop,
university medical database, television
program, business magazin, open-access
journal, blog, portal, daily newspaper,
scientific journal, online dictionary, trade
magazin, private supplementary health
insurance, manufacturer of fast e-bikes, ...

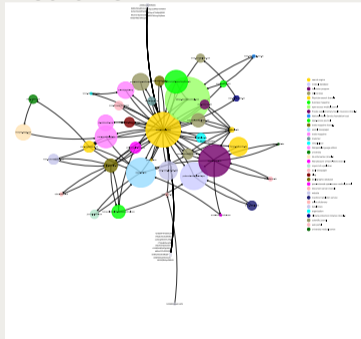
@2: SEARCH SPACES AND PATHS

color: type of website (Google is yellow), size: duration of visit, edge: link path

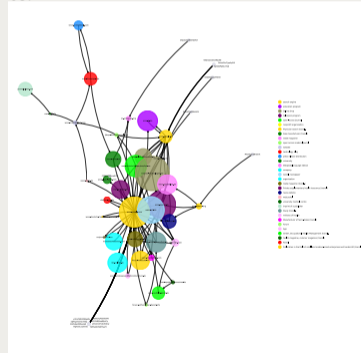
teacher trainee



medicine



law

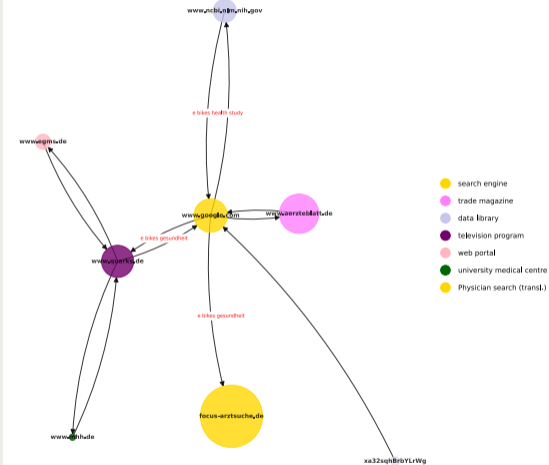
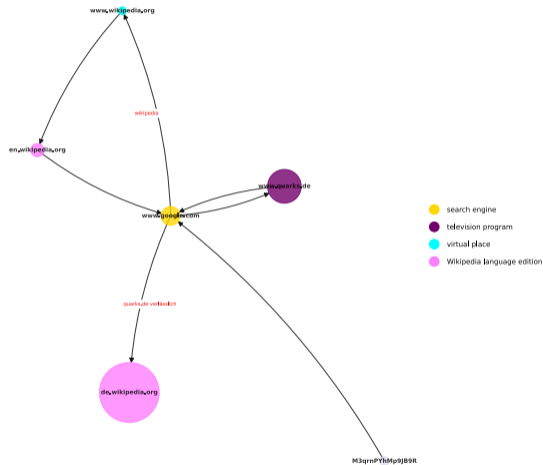


Of course, a search engine is the center of an online search.

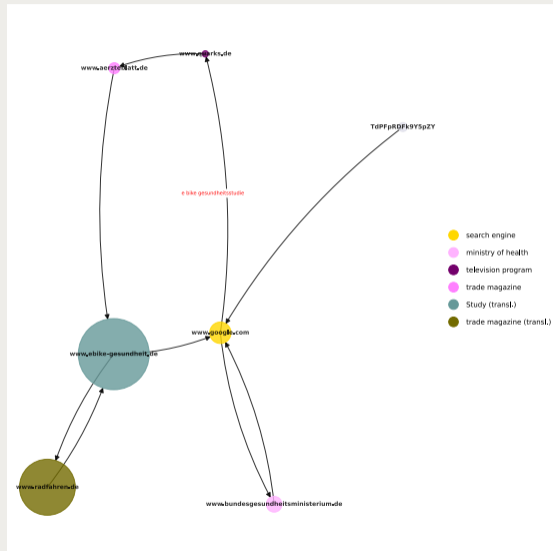
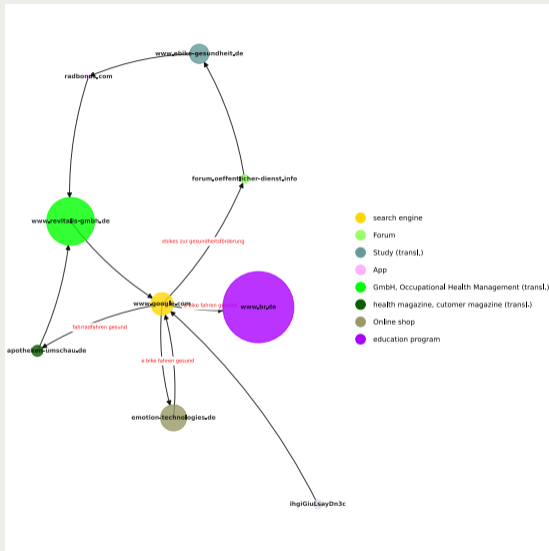
But two retrieval patterns become visible across domains:

- follow hyperlinks
- strictly toggle between search engine and retrieved websites

INDIVIDUAL EXAMPLES MEDICINE



INDIVIDUAL EXAMPLES LAW

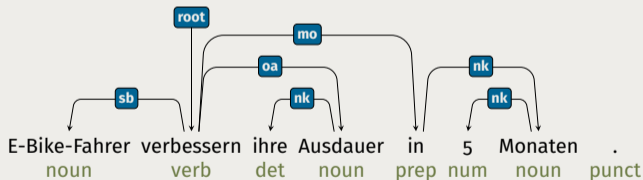


Main question: do the **websites differ in terms of text characteristics?**

A profile for each text/website is generated from a set of features.

- General features
- Lexical features
- Syntactic features
- Lexical features (HTML)
- Syntactic features (HTML)
- Lexical cohesion features

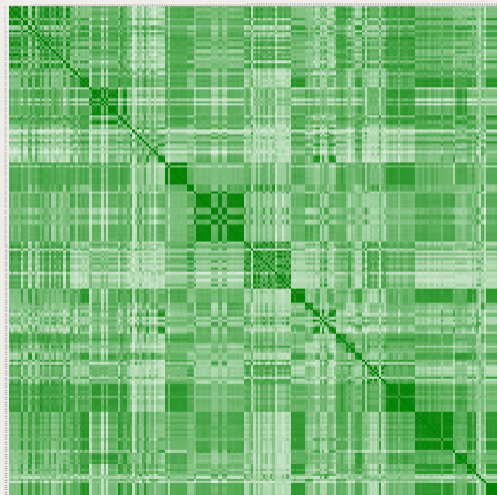
concrete features: count features (we see a more abstract one below)



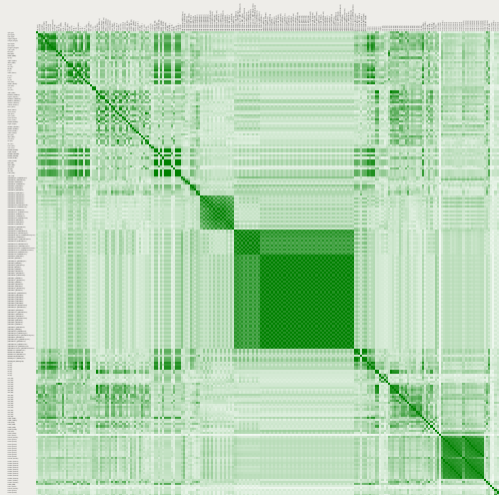
> 300 features in total

@2: TEXT SEPARABILITY AND SEPARABILITY FORCE OF FEATURES

text similarity



feature similarity (green square: HTML)



@2: NON-PARAMETRIC ANOVA: KRUSKAL-WALLIS H-TEST

The websites which are visited only by a single domain can indeed be separated by a number of features. The top ones include:

teacher vs. med (93 feat $p \leq 0.05$)

f34	POS	0.00000002
lmu	syn.	0.00000005
lG	syn.	0.00000018
bsesim	cohs.	0.00000063
lnHTMLbr	HTML	0.00000169
bcmu	syn.	0.00000218

law vs. teacher (57 feat $p \leq 0.05$)

G	STO	0.000391
ttr	STO	0.00166
f9	STO	0.00182
h	STO	0.00361
btgest	cohs.	0.00383

law vs. med

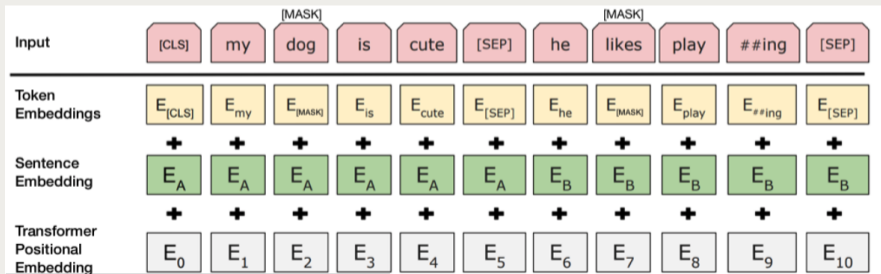
(108 feat $p \leq 0.05$)

lnHTMLbr	HTML	0.00000092
bsesim	cohs	0.00000096
Lmu	syn.	0.00000155
f34	POS	0.00000161
cG	syn.	0.00000321

@2: FEATURE CLASSES

among the best separating kinds of features are

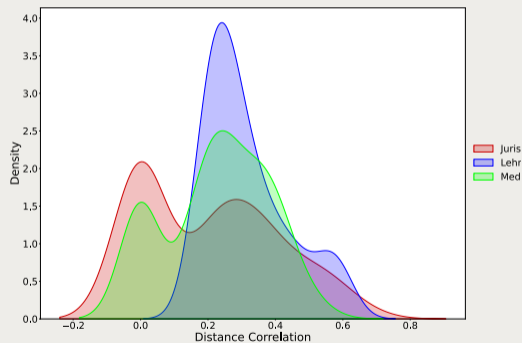
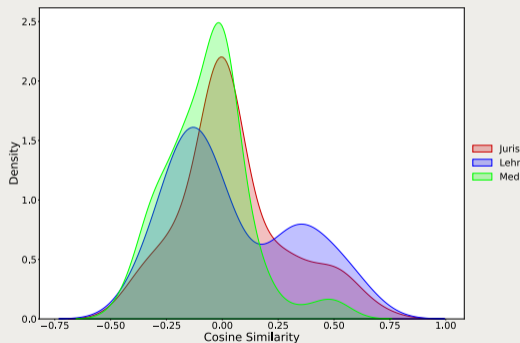
- Statistical text organization (STO; e.g. freqs., type–token ratio, entropy, ...)
- cohesion (BERT-based, see below)
- syntax (dependency-related)
- HTML



²J. Devlin et al. (2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: Proceedings of NACL, pp. 4171–4186.

@2: DISTANCE FROM INTERSECTION

Do the domains differ wrt. to the similarity of the websites visited exclusively to those visited by all?

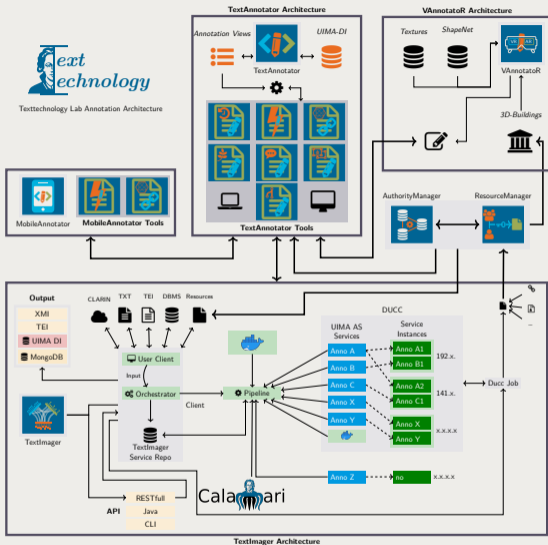


(Note that negative distance correlations are due to Gaussian smoothing)

- Q: critical online reasoning: domain-specific or general?
- domain differences in search behavior
- domain differences in quantitative linguistics profiles of consulted websites
- ➔ new computational linguistics method for assessing text characteristics in educational contexts
- ➔ follow-ups: correlate text characteristics and
 - ▶ task success / polarity
 - ▶ website credibility
 - ▶ domain prediction / website recommendation
 - ▶ ...



TextTechnology Lab Annotation Architecture



bridge.uni-mainz.de
www.plato.uni-mainz.de
www.texttechnologylab.org

LIST OF FEATURES

G, ttr, f9, h, btgest, f15, bcH, ccH, depH, LH, cH, imbH, WH, UG, SPEH, LDEH, hl, lnHTMLsup, lmbd, f3, lnHTMLq, f1, R1, HTMLccmu, HTMLbcmu, btgclmu, lnHTMLlol, btsimH, A, HTMLprmu, HTMLprH, HTMLccH, SPEG, lnnpd, npd, lnHTMLb, lnHTMLh6, preppd, lnHTMLblockquote, lnHTMLstyle, bt_first_lsH, bt_min_lsH, bt_max_lsH, btlsH, bt_mean_lsH, bt_prod_lsH, btgrc, lnHTMLh2, cG, bsest, bsclmu, RRR, bt_max_Ha, f2, bsccmu, bt_mean_Ha, lnHTMLp, lnHTMLmsub, lnHTMLmath, lnHTMLcaption, lnHTMLannotation, lnHTMLdl, lnHTMLmo, lnHTMLmrow, bscsdet, bt_min_Ha, bt_first_Ha, ppd, bt_prod_Ha, lnHTMLsub, btglclr, f16, btHa, lnHTMLspan, LDEmu, adjpd, lnadjpd, vpd, apd, btglclH, NDW, SPEmu, f31, f10, L, lnHTMLdt, btsH, LDEG, bcG, WG, btrac, lnHTMLbr, f4, alpha, lnHTMLh5, TCImu, VD, bt_prod_sH, lnHTMLsmall, TCIG, lnHTMLtr, btgass, Lmu, btdet, lnHTMLdiv, lnHTMLh4, advpd, ATL, bt_first_sH, bt_min_sH, f88, Q, MDDmu, lnHTMLmn, lnHTMLdd, lnHTMLmi, lnHTMLmark, LG, f27, f7, ccG, imbG, Wmu, f17, bt_max_rac, dpd, bt_mean_sH, cmu, bt_max_sH, ipd, lnHTMLarea, lnHTMLmstyle, lnHTMLsemantics, lnHTMLcode, bt_max_adc8, bt_prod_adc7, bt_first_adc3, bt_min_adc7, f35, f29, HTMLbcH, bt_prod_ac7, bt_mean_rac, bt_min_ac7, bt_first_ac3, bt_first_adc8, bt_prod_dfa, bt_min_h, bt_min_dfa, MDDG, bt_prod_adc3, bt_mean_lH, bt_min_lH, bt_prod_lH, bt_first_lH, bt_lH, bt_max_lH, bt_first_adc4, bt_max_ac8, f30, bt_max_ac1, bt_mean_dfa, bt_prod_adc4, lnHTMLlabel, bssimH, bt_min_adc3, lH, f37, f28, lnHTMLul, bt_mean_ac3, bt_first_dfa, bt_min_adc4, bt_mean_adc4, bt_mean_adc3, bt_first_adc7, bt_first_ac7, bsesim, bt_min_ac3, bt_prod_ac3, bt_first_ac8, bt_mean_adc6, bt_mean_adc8, bt_max_adc4, f6, f18, bt_max_adc1, H, btac7, bt_min_adc8, bt_prod_ac4, wH, btac6, bt_max_ac10, bt_first_ac4, f41, f32, ASL, tc, bt_max_adc7, pHTMLdl, lnHTMLhr, pHTMLstyle, pHTMLmo, pHTMLh4, pHTMLq, pHTMLinput, pHTMLmn, pHTMLarea, pHTMLmstyle, pHTMLlol, pHTMLmrow, pHTMLdiv, pHTMLsup, pHTMLp, pHTMLspan, pHTMLh6, pHTMLb, pHTMLbr, pHTMLtr, pHTMLul, pHTMLh2, btadc7, bt_min_ac4, btadc6, bt_min_adc6, btradc, bt_max_dfa, bt_max_h, wmu, bt_mean_ac6, bt_max_radc, bt_prod_H, bt_min_ac10, bt_min_ac6, bt_max_adc6, bt_prod_adc6, bt_first_ac10, bt_prod_adc8, bcmu, f33, f39, bt_prod_lHa, bt_lHa, bt_first_lHa, bt_mean_lHa, bt_min_lHa, bt_max_lHa, btadc4, bssimmu, bt_prod_ac10, stc, bt_first_adc6, bt_min_radc, btac10, btac4, lnHTMLh3, bt_first_ac6, bt_first_adc9, bt_first_radc, bt_mean_adc5, bt_mean_ac4, bt_max_ac6, bt_mean_ac8, btadc5, bt_mean_adc7, bt_min_ac8, btac9, bt_prod_radc, btadc8, bt_max_ac4, bt_mean_ac10, depmu, bt_min_H, bt_prod_ac6, bt_max_H, bsd, bt_mean_radc, bt_first_rac, lG, bt_mean_ac1, bt_prod_ac8, btac5, btac3, bt_min_adc5, bt_max_adc2, btac1, bt_max_ac3, ccmu, btsim, bt_max_adc3, btac8, bt_prod_adc5, bt_max_ac7, bt_mean_ac7, bt_mean_adc1, bt_max_ac2, bt_first_H, bt_first_adc5, btadc9, bt_min_ac2, bscch, charH, bt_prod_ac2, bt_max_adc5, bt_first_ac9, bt_prod_adc2, bth, btadc1, imbmu, bt_max_adc9, bt_min_adc2, lmu, TCiH, btdfa, bt_first_h, bt_min_ac5, bt_min_adc9, bt_mean_adc9, bt_prod_ac5, bt_mean_h, bt_prod_h, pHTMLli, pHTMLblockquote, pHTMLcaption, pHTMLmark, pHTMLsub, pHTMLmi, pHTMLdd, pHTMLh5, pHTMLh3, pHTMLlabel, pHTMLsmall, pHTMLimg, pHTMLdt, pHTMLth, pHTMLtable, pHTMLmsub, pHTMLi, pHTMLannotation, pHTMLa, pHTMLcite, pHTMLlink, pHTMLtd, bt_min_ac1, btadc3, lnHTMLinput, bt_mean_ac5, bt_prod_ac1, bt_prod_adc9, bt_min_adc1, bth, bt_max_ac5, btac2, bt_prod_adc1, MDDH, bt_min_ac9, bt_min_rac, bt_first_ac2, bt_mean_adc2, bt_first_adc2, bt_mean_H, bt_first_ac5, RR, bt_first_ac1, btgd, wG, btadc2, bt_prod_ac9, bt_max_ac9, bt_prod_rac, depG, bt_mean_ac2, f34, bt_mean_ac9, bt_first_adc1