Resolving code names to structures from the medicinal chemistry literature: Not as FAIR as it should be

Roxana-Maria Rujan ORCID 0000-0002-4335-3338

Data Science, Medicines Discovery Catapult,
Macclesfield, UK, SK10 4ZF https://md.catapult.org.uk/

Outline

- Example Compound Codenames (CN)
- CN in the wild
- Rogues' gallery
- Manual collation of CNs
- Automatic CN retrievals
- Curation of CN from J Med Chem
- FAIR-ness aspects
- Concluding remarks and plans

CN from two years of J.Med.Chem

S1-6e UZH2
RPE65 TAS-116 GEP44
TAK-981 MK-8262 SPR519 OI338
GC-14 MK-8153 FRM-024 LYS006 SLC-149 LU13
TAK-020 FGF401 CC-90001 BLU-945 LFS-829 FM26
TO-317 EC5026 UCB7362 S06-1011 CC-90005 NCS-382
ZW4864 CJ-2360 PF-562271 I-BET282E S11-1014 GAT211 K117
EX527 CX-4945 MRTX0902 FPFT-2216 GLPG2451 SHU-9119 MT-3995
MK-4688 S-217622 CHF-6366 BAY-6672 GDC-0276 PQQ GB1211 T-914
RP-6685 SPH3127 BPR1R024 LY3202626 BAY-8400 PRN1008 MG624 D24
SLL-627 ORG27569 VU6019650 CH7057288 MMV665917 MRS8054 JXL001 K-5a2
SK-575 NUC-1031 BAY-4931 BMS-986202 BMS-986313 CHDI-626 QCP01 ST-2001
AM11245 BAY1217224
SJ6986 SCO-267 LSN3318839 ACT-672125 BMS-820132 BCX7353
SY-5609 NEt-C343 CCF0058981 PF-06865571 ANAVEX2-73 AZD4831 TNIR7-1A 13b-K
LEI-401 GNE-064 BMS-963272 GSK3640254 SUVN-D4010 RGD-SS-CA AS-1763 HP590
Hu7691 EST73502 AzoGW1929 SGC-STK17B-1 IACS-15414 SPH5030 3a-P1
C20 INE963 EST64454 NCATS-SM1440 R)-AS-1 SAGE-718
TH257 CNP520 BRD0639 ACH-000143 IHMT-MST1-58 PF-00835231 PLHSpT LYC-55716 TK-129
RB394 CC-90009 LT-850-166 PF-06455943 QNX-sLXms GSK3739936 NSC791985 AS-0141 PF-74
CM-444 PSMA-1092 GSK3494245 N-CTX-Ltg1a
HN37 OATD-01 ASTX029 PF-06835919 CHNQD-01255 NTQ1062 BMS-986120 MRTX1133 UMB298
UCF501 CQ211 KSL-128114 BAY-1101042 S)-VU0637120 DC-PRC2in-01 PBI-4DNJ-1 QN523 V-0219
L24 CBTF-EE NVP-CLR457 GSK2982772 MRTX1719 WS-691
MP135 808-NM2 LY3154885 ARUK3001185 R)-STU104 QPX7831 BMS-986144 SR18292 XL-147
A17 BIIB091 AZD9833 JNJ-64264681 ACT-1004-1239 PF-06843195 QBW251
SZM679 GQ127 AZD0284 JNJ-63576253 Ga-bisNODAGA ABBV-3373 BGB-290 AG-270 K122
E197 PF-07059013 MSD-496486311 HY-1
MRIA9 VZMC013 ARD-2585 PF-06873600 NCATS-SM5637 ARN-21934 CRCM5484 GW4064
SHA-68 IPN60090 DDD01305143 BMS-986176 HST5040 MCoTI-II
23dd UNC6641 JDQ443 ACT-660602 JNJ-54861911 GS-441524 I-BET567 AF27139
CK-274 AZD8154 DNDI-6148 GNE-0749 MAK683
R)-7 BT8009 EEDi-5273 BMS-955176 SIAIS164018 BAY1214784 PQR620 I-BET432 KK-052 ZN-c3
AP1 KVD900 ICRF-193 BMS-986278 BMS-986180 GDC-2394 GSK251 XY123
M2698 INCB054828 BMS-986339 ANT3310 ITH15004 MLT-231 F3A
KT-531 LEO39652 AZD4573 IHMT-PI3KÎ AZD5305 MSC-4106
RRx-001 SMU-C80 DDO-2213 ARN19689 GLPG1972 DFX-PtÂ M-23
D6808 CVN424 LP-922056 DC-BPi-07 GLPG1205 UCM-1306 RV521
A23 NB-360 BAY-091 IDR-1002 HSGN-218 SGC6870 LX-9211
S64315 BAY-069 SLL-1206 NIC-0102 BI-3406 PI-2620 TAM16
ZCM-I-1 KGOP01 CC-90011 TNP-2198 GSK452 A1-31
XD2-149 MK-1454 heMAMP CJ2-150 MK-0159 SW-101
WSJ-557 RP-6306 RLX-33 TNO155 ID33
P18 UK-5099 TUG-891 B53
ID09

CN in the wild : good and bad news

- Full codename-to-structure (CNn2s) with SAR in quality journals defines the leading edge of global drug discovery
- “XXX-12345” is predominant
- Many leads only have codeless locants from the papers (e.g., “compound 21”)
- It would be good if authors performed specificity checks and homonym clashes (e.g., Google, PubMed, PubChem) – e.g., A17 compound
- Can get CN “daisy chains” via company mergers and acquisitions
- Companies re-badging “old chestnut” compounds with new CN

CN in the wild : good and bad news

- Guide to Pharmacology, BindingDB and ChEMBL annotate many CN2s and submit to PubChem
- Curation is selective and lags by 2- 12 months
- In PubChem multiple CN + INN + USANs + trade names leads to "synonym spaghetti"
- Confounded by 45% of USANs being mixtures
- Same CN systems for biologicals as well as small-molecules
- Press release CN and majority of Phase I and II clinical trial compounds are still "blinded" (no CN2s)

Rogues Gallery: confounding “Findability” in FAIR

- Overdoing the digits most promising molecule, 12126065, exhibited antiviral > 8 spurious matches in NCBI databases
- Confusing and search-challenging punctuation **(*R*)-7** [**(*R*)-AS-1**] showed broad-spectrum antiseizure activity
- The great hyphen ambiguity > 3-fold permutation across data sources

American Chemical Society
https://pubs.acs.org › doi
AZD0284, a Potent, Selective, and Orally Bioavailable Inverse ...

MedChemExpress
https://www.medchemexpress.com › AZD-0284
AZD-0284 | RORγ Inverse Agonist

Springer
https://adisinsight.springer.com › drugs
AZD 0284 - AdisInsight - Springer
2 Jun 2022 — **AZD 0284** was a selective inverse agonist c

- Chinese teams using alphabetic locants that look like CN Mechanistically, **L24** exhibited significant *in vivo* antitumor effects

Rogues Gallery: confounding “Findability” in FAIR

- Peptide challenges

  with toxins from a scorpion and a spider. Here, we described a cone snail toxin, N-CTX-Ltg1a, with a

- Zentalis Pharma – needs an internal registration system e.g. ZNPxxxxxx?

  of compound **16** provided an unexpected improvement of Wee1 potency. Compound **16**, known as ZN-c3, showed excellent *in vivo* efficacy and is currently being evaluated in phase 2 clinical trials.

- Shanghai Institute for Advanced Immunochemical Studies – shorten to SIAxxxxxx?

  Discovery of a Brigatinib Degrader SIAIS164018 with

MDC internal Codename project - a competitive intelligence initiative

- Journal of Medicinal Chemistry PubMed Abstracts were curated for the detection of new CNs potentially resolvable to lead molecules
- This harvests latest drug discovery outputs , including AI/ML designed compounds
- Resolved compounds can now be explored *in silico* by MDC informatics and sourced for *in vitro* experiments by MDC Discovery Sciences
- Our internal feed is months ahead of other sources
- From quality journals, manual curation of the latest CNs is possible

REGEX > automated CN mining

- Automated recognition of CNs from PubMed abstracts or any text
- Used 300 CNs as starting point but tune to include new ones
- Compiled False Positive (FP) blacklist including gene names and cell lines
- Used PubMed counts as a True Positive (TP) filter
- Short CNs still challenging for FPs

REGULAR EXPRESSION

580 matches (48 444 steps, 8.7ms)

```
/ (^|\s|\(|\[|\])(\w{5}\d{4}|[A-Z]{3}\d|[A-Z]\d{6}|[A-Z]\d{2,4}|[A-Z]{2,5}\d{2,7}|[A-Z]{4}\-
\d{2}|[A-Z][a-z]\d{4}|[A-Z]\d\-\d[a-z]|[A-Z](\-|\s)\d{3,6}|[A-Z]{2,6}\d\-\d{2,3}|[A-Z]
{4}\-[A-Z]{2}|[A-Z]\d(\-|\s)\d{2}|[A-Z]{2,5}(\-|\s)\d{2,9}|[A-Z]{3}\d[A-Z]\d{3}|[A-Z](\-
|\s)\d[a-z]\d|[A-Z]\d{2}(\-|\s)\d{4}|[A-Z]{2}(\-|\s)[a-z]\d|\d{3}(\-|\s)[A-Z]{2}\d|[A-Z]
(\-|\s)[A-Z]{3}\d{3}|[A-Z]{2}[a-z](\-|\s)\d{3}|[A-Z]{3}[a-z](\-|\s)\d{4}|[A-Z]{4}\-\w{4}
[α-ω]\-\d{3}|[A-Z]{3}\-[A-Z]{2}\-[A-Z]{2}|[A-Z]{3}(\-|\s)[A-Z]{3}\d{3}|\([A-Z]\)(\-|\s)[A-
Z]{2}\d{7}|\([A-Z]\)(\-|\s)[A-Z]{3}\d{3}|[a-z]{2}[A-Z]{4}|[A-Z]{3}\d{8}|[A-Z](\-|\s)[A-Z]
{3}\d{3}[A-Z]|[A-Z]{3}(\-|\s)\d{5}|[A-Z]\d{5}|[A-Z]{3}\-[A-Z]{3}\d{2}[A-Z]\-\d|[A-Z]{2}(\-
|\s)\d{3}(\-|\s)\d{3}|[A-Z]{4}[a-z][A-Z]|[aA-zZ]{2}(\-|\s)[a-z]{3}[A-Z]{6}|[A-Z]{3}(\-
|\s)\d[A-Z]{3}(\-|\s)\d|[A-Z]{3}(\-|\s)[a-z][A-Z]{2}[a-z]{2}|[A-Z]{4}(\-|\s)[A-Z]{3}\d(\-
|\s)\d{2}|[A-Z]{2}(\-|\s)[A-Z]{2}[a-z](\-|\s)\d{2}|[A-Z](\-|\s)[A-Z]{3}(\-|\s)[aA-zZ]
```

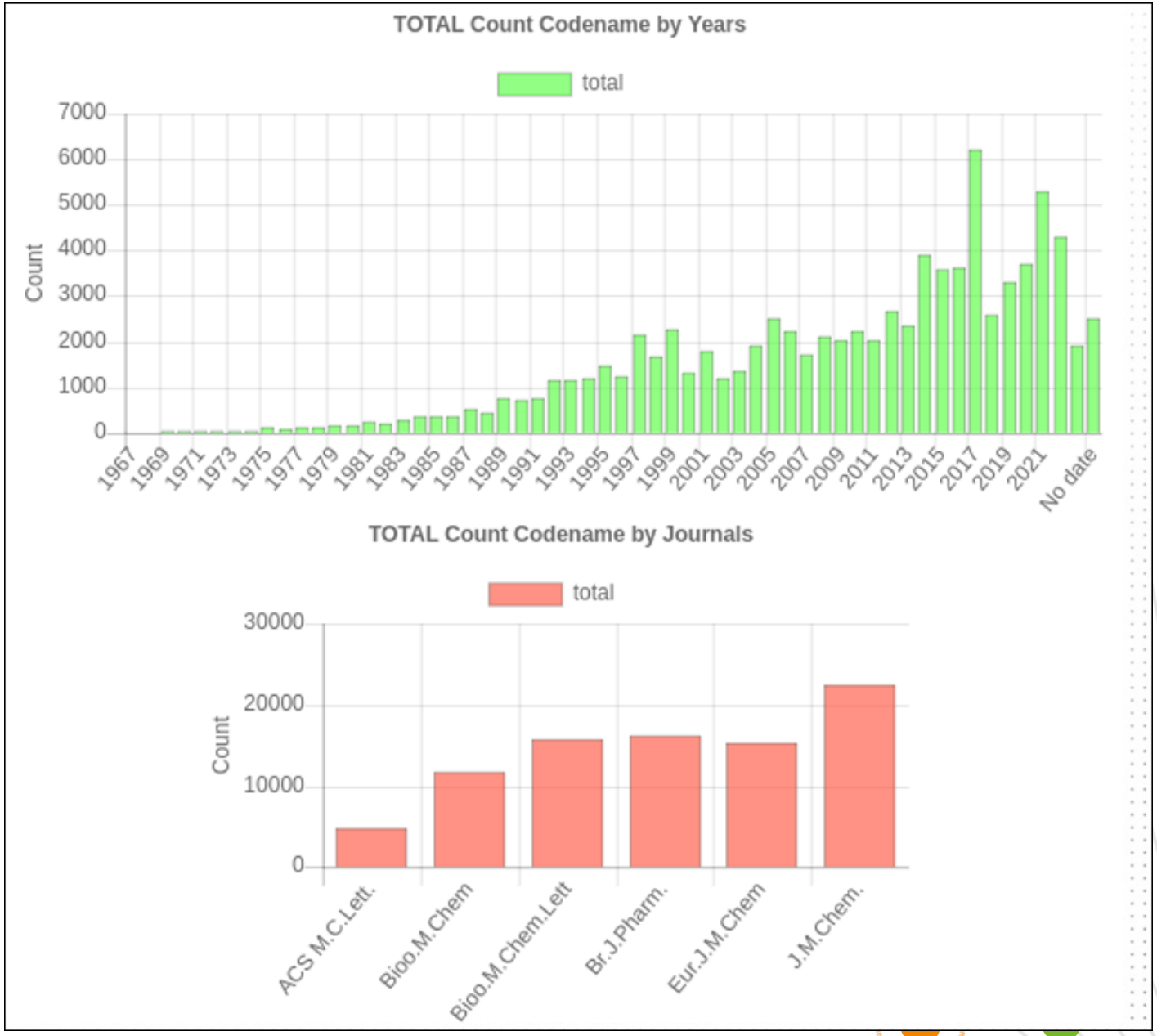
/ gm

TEST STRING

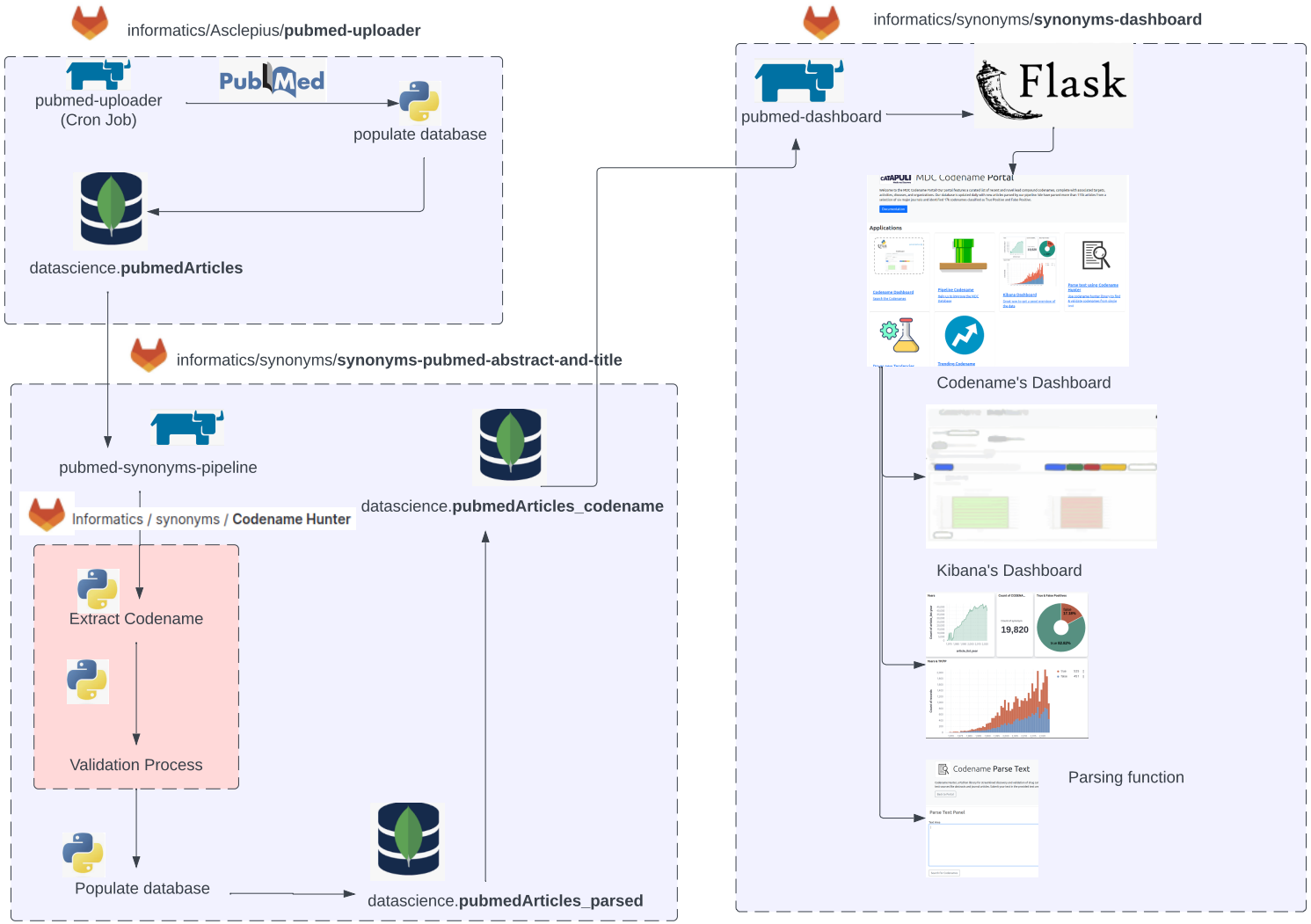
```
##########################################################################################
#########################################Codenames#################################################

 LU13  [18F]LU13  S07-2010  NDT-30805  K11777  SPR7  EZM0414  CHNQD-01255  ORG27569  CH7057288  NIC-
0102  N-CTX-Ltg1a  UCM-1306  GSK3640254  BMS-955176  IHMT-MST1-58  AZD4831  (R)-7-[(R)-AS-1]  ACT-
672125  ACT-660602  S11-1014  S11-1033  SPH3127  GNE-064  SLL-627  RP-6306  BLU-945  MG624  MRTX0902  
GW4064  SAGE-718  LU13  MSC-4106  MK-0159  BMS-986120  BMS-986141  BMS-986339  NTQ1062  MT-3995  RLX-
33  CQ211  NVP-CLR457  (R)-STU104  QN523  VU6019650  S-217622  V-0219  CRCM5484  MK-1454  SPH5030  
BMS-986176  LX-9211  GSK3739936  BMS-986180  BMS-820132  LY3154885  TNP-2198  MRTX1719  I-BET567  
FPFT-2216  TAM16  CC-90001  MRTX1133  QPX7831  BIIB091  SY-5609  MK-4688  DNDI-6148  BMS-986278  BAY-
091  CNP520  IACS-15414  BMS-963272  EEDi-5273  AS-0141  BPR1R024  CHNQD-01255  ORG27569  CH7057288  
NIC-0102  N-CTX-Ltg1a  UCM-1306  GSK3640254  BMS-955176  IHMT-MST1-58  AZD4831  (R)-7-[(R)-AS-1]  ACT-
672125  ACT-660602  S11-1014  SPH3127  GNE-064  SLL-627  RP-6306  BLU-945  MG624  MRTX0902  GW4064  
SAGE-718  LU13  MSC-4106  MK-0159  BMS-986120  BMS-986339  NTQ1062  MT-3995  RLX-33  CQ211  NVP-CLR457
  (R)-STU104  QN523  VU6019650  S-217622  V-0219  CRCM5484  MK-1454  SPH5030  BMS-986176  LX-9211  
GSK3739936  BMS-986180  BMS-820132  LY3154885  TNP-2198  MRTX1719  I-BET567  FPFT-2216  TAM16  CC-
90001  compound 13e  Compound 13e  Compound-13e  Compound-13E  compound-13e  compound-13E  compound 
347  compound 64716  T-00127-HEV2  SHM115  CX-4945  BGB-8035  BGB-3111  RK-701  OY-101  ACT-777991  
ZZ151  RG7907  AZD5305  FRM-024  LT-850-166  AS-1763  GSK251  LYC-55716  ARD-2585  ARN19689  AZD0284  
PI-2620  TAK-020  BCX7353  UZH2  BAY-8400  T-914  ZN-c3  ASTX029  SLL-1206  I-BET282E  ZW4864  Hu7691  
 MK-8262  DC-BPi-07  CC-90005  CHDI-626  CCF0058981  MMV665917  BRD0639  SPH3127  GNE-064  SLL-627 
```

Automated CN mining from key Journals

- ~ 19K PubMed CNs from 6 journals
- J.Med.Chem has most CNs
- Includes some false positives (FPs) but frequency is low
- Some CNs not in abstract
- Provisional portfolio mappings
  - PF- = 142
  - GSK = 136
  - AZD = 93



Codename Pipeline



EXCELLENCE INTEGRITY COMMUNITY INNOVATION

Codename Pipeline – Codename Dashboard

Codename Dashboard

Introducing our Codename Dashboard: the ultimate solution for pharmaceutical compound identification. Uncover codenames, analyze journal frequencies, and explore publication years with ease. Seamlessly navigate the dynamic landscape of drug development. With daily updates, our pipeline ensures real-time data availability for your critical research. Immerse yourself in the power of our robust and up-to-date dashboard.

Back to Portal

Search Panel

Codename:

Examples of codename

submit

Difficulty finding the codename?

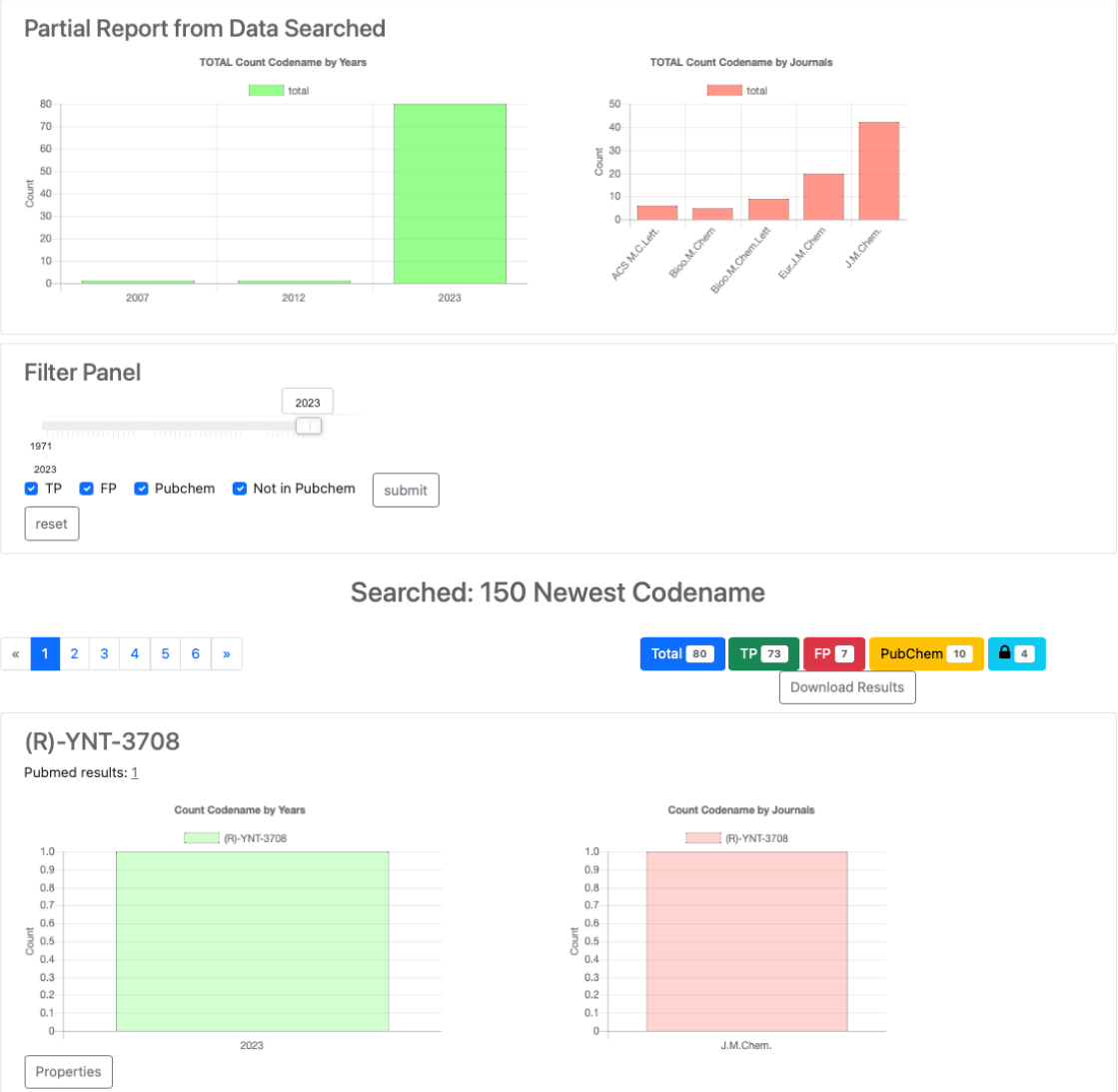
Upload a list of PubMed IDs

Choose File no file selected

Database Statistics

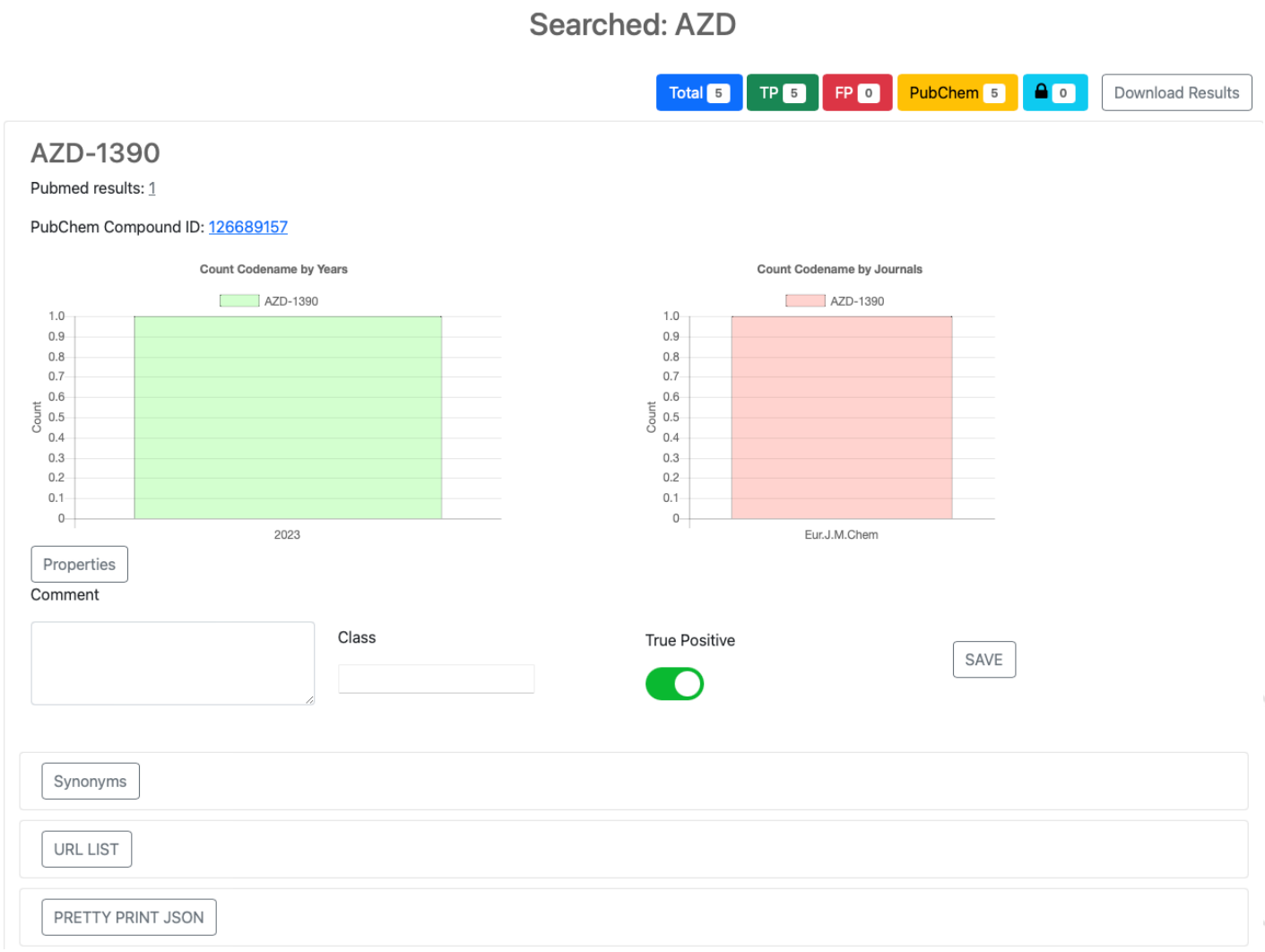
See more Information

Codename Pipeline – Codename Dashboard



EXCELLENCE INTEGRITY COMMUNITY INNOVATION

Codename Pipeline – Codename Dashboard



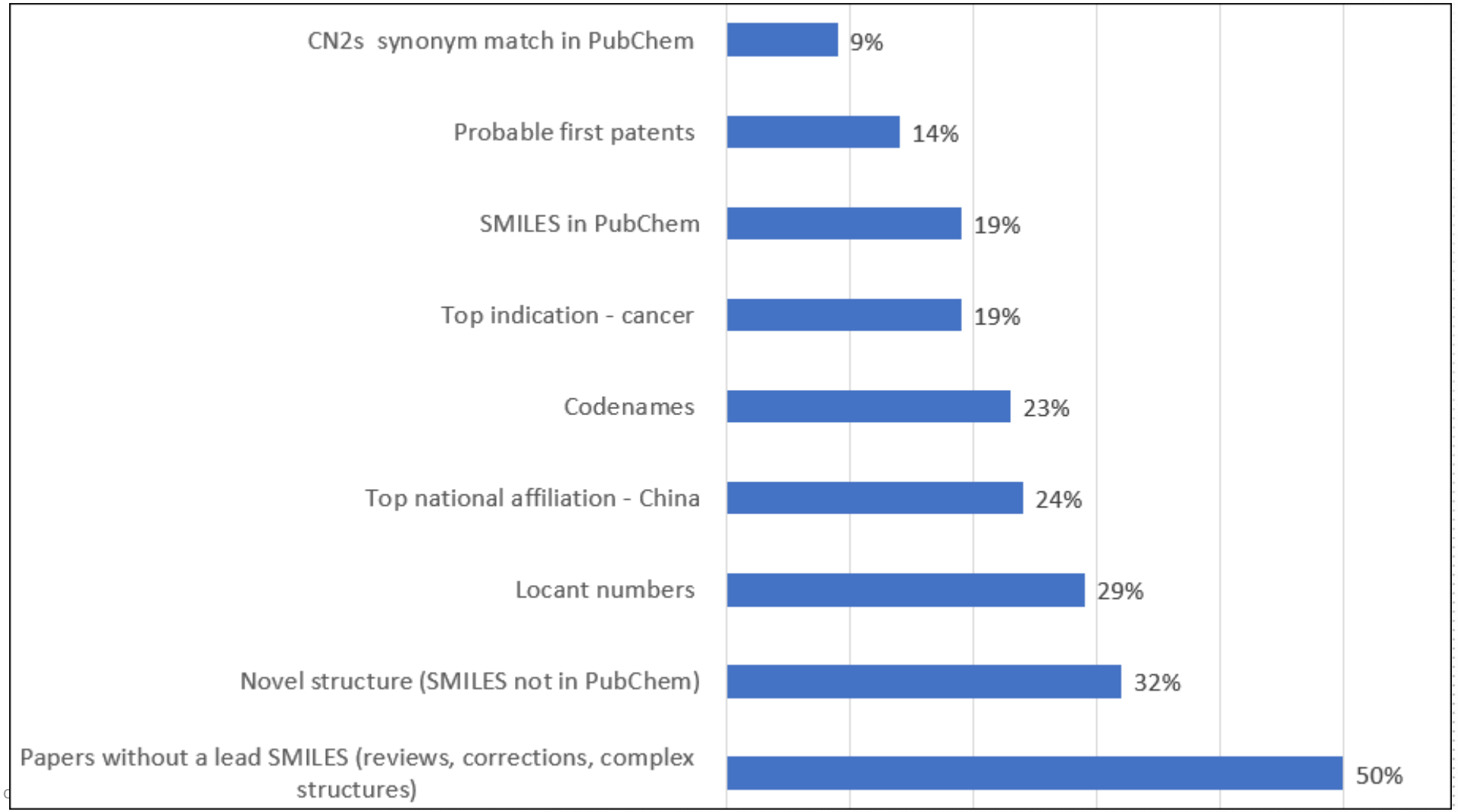
Manual curation results from 150 most recent J.Med.Chem

PMID	Title	First Author	PMID date	Codename	Lead name or exclu	Synonym	InChIKey	PubChem CID	PC CN2s
37145846	X-ray Crystal Structure-Guided Discovery of N	Ren Y	05/05/2023		**cpd 3a**				
37145039	Rationally Engineered CYP3A4 Fluorogenic Su	He RJ	05/05/2023		substrate, no SMILES				
37141440	Discovery of Potent and Wild-Type-Sparing Fc	Dong H	04/05/2023		**cpd D51**		YETCUDGHQDLIRZ-UHFFFAOYSA-N	163288439	N
37140467	Prodrug Strategies for the Development of β-	Singh US	04/05/2023	**ODE-I-BHDU-MP**		cpd 38			
37134237	Design, Synthesis, Biological Evaluation, and (	Skácel J	03/05/2023		**cpd 45i**		IMXBQYQAJSJWFZ-ONEGZZNKSA-N	156316535	N
37134203	Design, Synthesis, and Evaluation of (R)-8-((T	Li Q	03/05/2023		**cpd 10zi**				N
37134182	Small Molecules Targeting DNA Polymerase T	Pismataro MC	03/05/2023		Review				
37133930	Virtual Special Issue: New Drug Modalities in	Aubé J	03/05/2023		Review				
37133411	Ψ and χ Angle Constrains at the C-Terminus T	Zhou Y	03/05/2023		**Peptide 3**				N
37130350	Discovery of a Potent and Oral Available Com	He P	02/05/2023	**HP661**					N
37130343	N-Phenyl-1-(phenylsulfonyl)-1H-1,2,4-triazol-	Lane T	02/05/2023	**12126065**					N
37130331	Design, Synthesis, and Bioevaluation of Nove	Chen P	02/05/2023		**cpd c17**		LDKDGYNCHLGOIP-UHFFFAOYSA-N	167440207	N
37130057	Fragtory: Pharmacophore-Focused Design, Sy	Bührmann M	02/05/2023		Library deisgn				
37130037	Transformation of a Dopamine D(2) Receptor	Liu R	02/05/2023		**cpd 29c**				
37129317	Discovery of Potent Tetrazole Free Fatty Acid	Valentini A	02/05/2023	**TUG-2304**		cpd 16l			N
37116172	Thiophene Carboxamide Analogs with Long A	Ohta K	28/04/2023		**cpd 2k**				

SMILEs if no PubCher	Target HGNC	arget UniPro	Disease indication	Disease Ontology	Mechanism	Activity
COC1=C(OC)C(OC)=CC(C2=NC=CC3=C2N=C(C4=CC=CC5=C4C=CN5			cancer	DOID_162	tubilin bindig	MCF-7 cells IC50 = 6nM
	EGFR	P00533	cancer	DOID_164	inhibition	IC50 = 14 nM (vs mutatant)
Br/C=C/C1=CN([C@H]2O[C@@H](COP(O)(OCCOCCCCCCCCCCCC			Varicella zoster virus (VZV)	DOID_8536	DNA synthesis inhibition	
	PNP	P00491	autoimmune	DOID_417	inhibition	1C50 = 52 nM
O=C1N(C[C@@H]2OCCC2)C	TNK2	Q07912	NSCLC (lung cancer)	DOID_0080521	inhibition	IC50 = 2.1 nM
O=C1N([H])CC(N2)=C(C[C@	MC4R	P32245	cardiovascular	DOID_1287	agonist	EC50 = 4.1 nM
O=C(N1CCN(CC2=CC=C(OCC(F)(F)F)C=C2)CC1)C3=C(C)N(CC4=CN			cancer	DOID_164	inhibitor	
N#C/C=C/c1ccc2c(S(=O)(=O)n3nc(Nc4ccc(C#N)c(		P04585	HIV-associated neurocognitive disorders		antiviral RT inhibito	
	MYD88	Q99836	Acute Lung Injury - inflamation	SYMP_0000061	downregulating cyt	
COC1=CC=CC2=C1OC3=C2C(	DRD2	P14416	antipsychotic	DOID_2468	partial agonist	
O=C(N[C@H](CC1=NN=NN1	FFA2	O15552	metabolic and inflammatory diseases	DOID_0060158	antagonist	
O=C(C1=CSC=C1)NCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCC			cancer	DOID_164	inhibitory for gliobl	
ClC1=CC=C2[C@@](CCCC2=	MCL1	Q07820	relapsed or refractory malignancies	DOID_164	inhibition	

Main Afiliation	Likely first patent	Notes
Southern Medical University, Guangzhou		likely colchecine binding site
China Pharmaceutical University, Chongqing	CN114380806	Osimertinib resistance mutant specific L858R/T790M/C797S
University of Georgia		
Czech Academy of Sciences	WO2021083438	Series of inhbitors also active against
Jinan University		Analogue of ASK120067 and osimertinib
University of Arizona		Similar to Melanotan II CID92432
East China Normal University		Targeting OXPHOS complex I, related to IACS-010759 CID86711931
Collaborations Pharmaceuticals		Target is within polyprotein, long code has specifity problems
Wenzhou Medical University	CN115710251	SPR analysis of MyD88 binding affects TLR4 interaction
ShanghaiTech University		
University of Copenhagen		While series given TUG-codes
Kyoto Pharmaceutical University		inhibiting mitochondri synergisitcally with temozolomide (TMZ)
Janssen		Outside ROF but good oral bioavailability

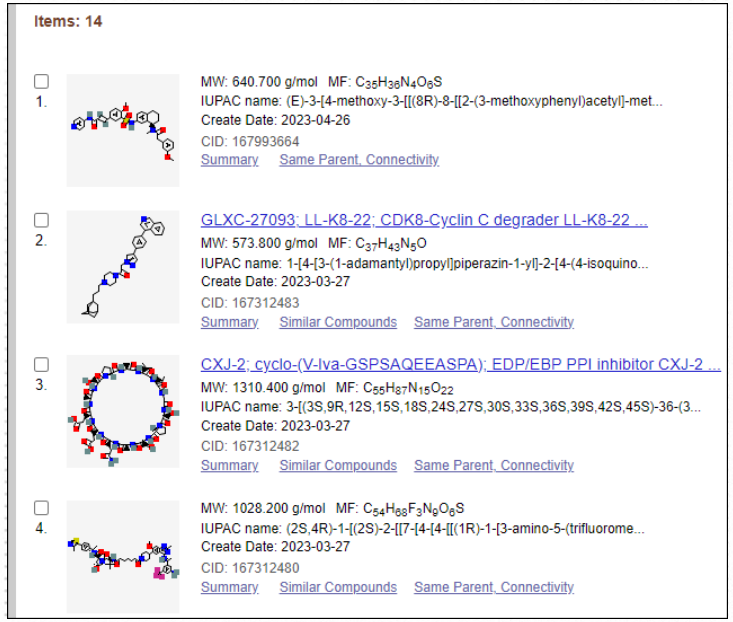
J.Med.Chem. manual curation -150: top-level stats



Most recent CN from J.Med.Chem

- From 33 curated CNs with a structure
- 14 had synonym match in PubChem (i.e. CN2s)
- 2 were in ChEMBL
- 2 in BindingDB
- 4 in Guide to Pharmacology
- Thus our codename curation from J.Med.Chem. gives MDC ~50% preview

BGB-8035
CCT374705
HUM-218 JG-2016
SLB1122168 LUZ5
HP661 SZM-1209 V2043
JSD26 MYF-03-176 RG7907
m-Se3 ODE-I-BHDU-MP CXJ-2
NXP800 R)-YNT-3708 IPG7336
TUG-2304 18F]NT160 LL-K8-22
SYNTiACT-777991 RK-701
OY-101 ARN24928 ZZ151
T-00127-HEV2
OICR-8268 CX-4945
CPD-1224
BAY-6096



FAIR solution to the locant problem (Guide to Pharmacology > PubChem)

Ligand: compound 21 [Zhang *et al.*, 2021]

Comments: **Compound 21** is one of two potential lead SARS-CoV-2 antivirals from the same
Bioactivity comments: **Compound 21** inhibits SARS-CoV-2 replication with an EC50 of ~1 muM . In the same

Ligand: compound 21 [PMID: 23981033]

Comments: **Compound 21** is reported as aSpirolactam-based lead **compound** inhibitor of acetyl-CoA carboxylase (ACC), inhibiting

Ligand: compound 21 [PMID: 24900635]

Comments: **Compound 21** is a derivative synthesised and assessed in a medicinal chemistry study to identify
Immuno Ligand Comments: **Compound 21** is a RIPK1 inhibitor with potential in diseases associated with necroptosis (programmed i

Ligand: compound 21 [PMID: 22802221]

Comments: **Compound 21** is a first-in-class orally administered angiotensin AT2 receptor agonist. It prevents
Immuno Ligand Comments: **Compound 21** is an adenosine AT2 receptor agonist that exhibits vascular anti-inflammatory effects in vi

Ligand: compound 21 [PMID: 34855405]

Comments: **Compound 21** is a small molecule apelin receptor agonist .
Bioactivity comments: **Compound 21** induces efficacy in a rodent heart failure model, in keeping with its molecular

Ligand: compound 21 [PMID: 23312943]

Comments: **Compound 21** is a potent inhibitor of TYRO3 protein tyrosine kinase (Sky kinase) . It is suitable
Bioactivity comments: **compound 21**, with an IC50 approximately 25 times less effective as compared to Sky inhibition

Ligand: compound 21 [PMID: 34141085]

Comments: **Compound 21** is an orally bioavailable tankyrase inhibitor that was designed for antitumour potential . It is active

CN2s curation: FAIR challenges

- Most Med Chem Journals are entombed in PDFs behind paywalls
- For JMC only ~ 12% of papers have CNs, with most specified by locants (e.g., 12 b)
- JMC supplementary .csv SMILES files high value – however
  - short on metadata
  - difficult to machine-read
  - may not include the CN from the text
  - activity result units and error (+/-) ranges get mangled in formatting
  - challenging to parse from Figshare
- Mapping CN2s is more difficult for other Med Chem Journals
- Need OPSIN for IUPAC > structure and DECIMER for image > structure
- Markush R-group nesting makes SAR extraction difficult
- Precision of both protein target naming and diseases mapping can be variable (same old story….)

Concluding remarks and plans

- MDC Codename project is feeding the latest curated J Med Chem CN2s to MDC Informatics, Discovery Scientists and project partners where appropriate
- Manual curation includes targets, bioactivity, disease, affiliation, PubChem CIDs, novel SMILES, molecular mechanism of action and patent mappings
- We intend to make the results open > subsumed into other public databases
- We are adapting our CN automated extraction to:
  - Apply to all recent PubMed abstracts
  - Extend portfolio mapping
  - Mine clinicaltrials.gov via disease selects
  - Intersect PubMed with PubChem searches
- Planning to investigate automated relationship extraction and contextual AI/NLP recognition of CN and compound numbers

Acknowledgments

MDC colleagues:

Christopher Southan ORCID 0000-0001-9580-0446
Miguel Amaral ORCID 0000-0002-8836-2617
Ian Dunlop ORCID 0000-0001-7066-3350

External:

- Prof Michael Gilson (PI of BindingDB) for instigation of author-specified SMILES as J.Med.Chem supplementary data
- Open resources for CN2s curation feeds