ChemDT: a deep learning framework for detection and translation of chemical names from scanned documents

Ruochi Zhang[1], Yajuan Huang[2], Liming Guo[2], Qiong Zhou[2], Qian Yang[2], Yan Wang[1,2], Kewei Li[2], Lan Huang[2], Yusi Fan[2,*], Fengfeng Zhou[2,3,*].

1 School of Artificial Intelligence, and Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, Jilin, China, 130012.

2 College of computer Science and Technology, and Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, Jilin, China, 130012.

3 School of Biology and Engineering, Guizhou Medical University, Guiyang 550025, Guizhou, China.

* Correspondence may be addressed to Fengfeng Zhou: FengfengZhou@gmail.com or ffzhou@jlu.edu.cn. Correspondence may also be addressed to Yusi Fan: fan_yusi@163.com.

Background and Aims of This Study

With the advent of artificial intelligence (AI) in recent years, drug discovery has witnessed a transformative shift (Fleming, 2018). AI significantly abbreviates the conventional drug development timeline from 4-5 years for drug candidate screening to a mere 8 months (Savage, 2021). AI algorithms are now extensively utilized across the spectrum of drug discovery, spanning from target identification (Schenone, et al., 2013), to lead compound discovery (Zhavoronkov, et al., 2019) , optimization(Tan, et al., 2021), and clinical trials (Patton, et al., 2021). However, these AI methodologies heavily reply on vast volumes of high-quality data for model training (Holzinger, et al., 2019). Therefore, efficient data acquisition has emerged as a critical challenge and a focal research topic in the AI pharmaceutical arena.

A central idea of small-molecular drug design is the understanding of the nexus between a molecule's structural attributes and its biological activities, facilitating the proposal of potential candidates (Savage, 2021). The vast reservoir of high-quality data have been archived in public and industry databases over the past decades (Gaulton, et al., 2012), which facilitate the development of diverse prediction models to elucidate the physiochemical mechanisms underlying small molecule behaviors, including the AI driven models for quantitative structure-activity relationship(QSAR) (Ghasemi, et al., 2018), molecular generation model (Jin, et al., 2018), and compound protein interaction (Tsubaki, et al., 2019). A salient feature rendering small molecules amendable to AI techniques is their structured molecular representation, typically via MOL-file (Dalby, et al., 1992), SMILES (simplified molecular input line entry specification) (Weininger, 1988), InChI (international chemical identifier) (Heller, et al., 2015).

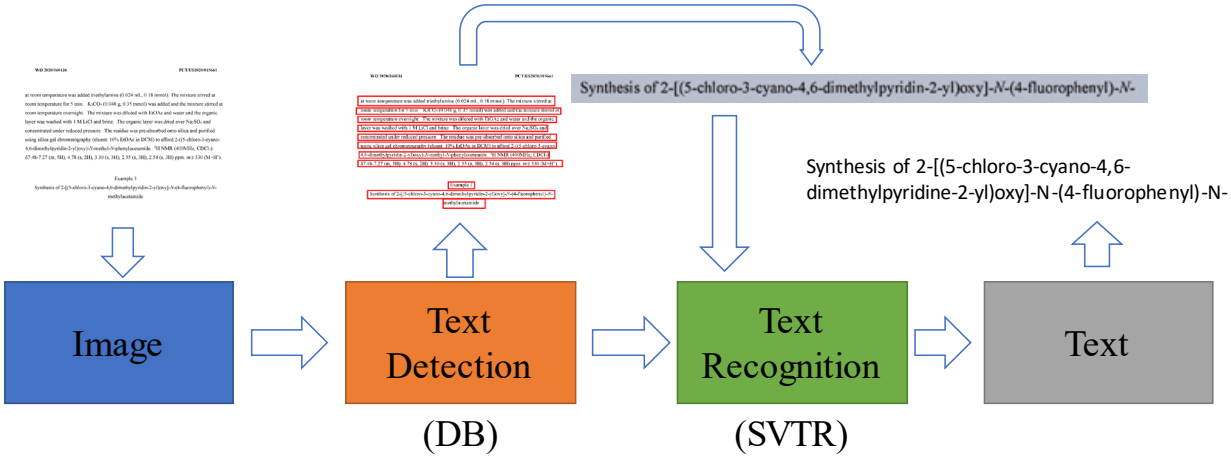
A substantial corpus of small molecule data exists in the scanned or printed formats as chemical structures or IUPAC names, and cannot be easily accessible to the AI model training. Several optical chemical structure recognition (OCSR) tools have been developed to automate the extraction of chemical structures from images in the literature (Rajan, et al., 2020) (Yoo, et al., 2022) (Oldenhof, et al., 2020). However, the task of IUPAC name recognition and its translation into structural formats remains relatively untapped. Usié, et al. introduced a linear conditional random field (CRF) (Lafferty, et al., 2001) for pinpointing IUPAC entities in the texts as of 2008 (Usié, et al., 2014). ChemSpot employed a similar approach using CRF combined with a dictionary (Rocktäschel, et al., 2012). A pivotal assumption underlying these tools is

the flawless character recognitions in images. However, the imperfect character recognition in real-world scenarios often leads to many false negatives in the IUPAC name recognition. The recent advances in name entity recognition (NER) methodologies have seen considerable enhancements (Li, et al., 2020), which may help improve the IUPAC name recognitions.

In light of these challenges and technological advancements, this study present ChemDT, a deep learning framework tailored for the detection of IUPAC chemical mentions in scanned or printed texts, and further translation of these names into molecular structures. We conducted a study on the disease target genes with the most citations in 2021 and downloaded the patent documents related to small molecule drugs associated with these targets. Subsequently, we employed the presented ChemDT software to extract the IUPAC names from these patent documents and converted them into molecular structural formulas. This data was then compiled to create the ChemDTD database. Both ChemDT and ChemDTD are open-source and freely accessible for non-commercial endeavors. ChemDTD further offers a streamlined graphical user interface, with regular updates on relevant targets and annotations. Our aspiration is for ChemDT and ChemDTD to empower researchers with accelerated data accumulation, and to ultimately expedite the drug discovery trajectory.

Materials and Methods

Optical Character Recognition



Supplementary Figure S1: Workflow of Optical Character Recognition Pipeline. The figure illustrates the end-to-end process, beginning with scanned or printed images containing text as input and culminating in the extraction of text strings as output. The text detection phase employs the Differentiable Binarization (DB) algorithm, while the Sequence-to-Sequence Visual Text Recognition (SVTR) algorithm is utilized for the text recognition phase.

The initial step in our workflow involves the identification of relevant text from scanned or printed images, a task commonly referred to as Optical Character Recognition (OCR). While various established OCR tools like Tesseract (Smith, 2007) and PaddleOCR (Du, et al., 2020) exist, they are generally ill-suited for recognizing IUPAC names with high accuracy. Conventional OCR systems struggle with the unique challenges presented by IUPAC strings, which can vary in length from tens to hundreds of characters. A single character misrecognition can invalidate the subsequent conversion to molecular structures, and this challenge underscores the need for an OCR model specifically optimized for IUPAC name recognition.

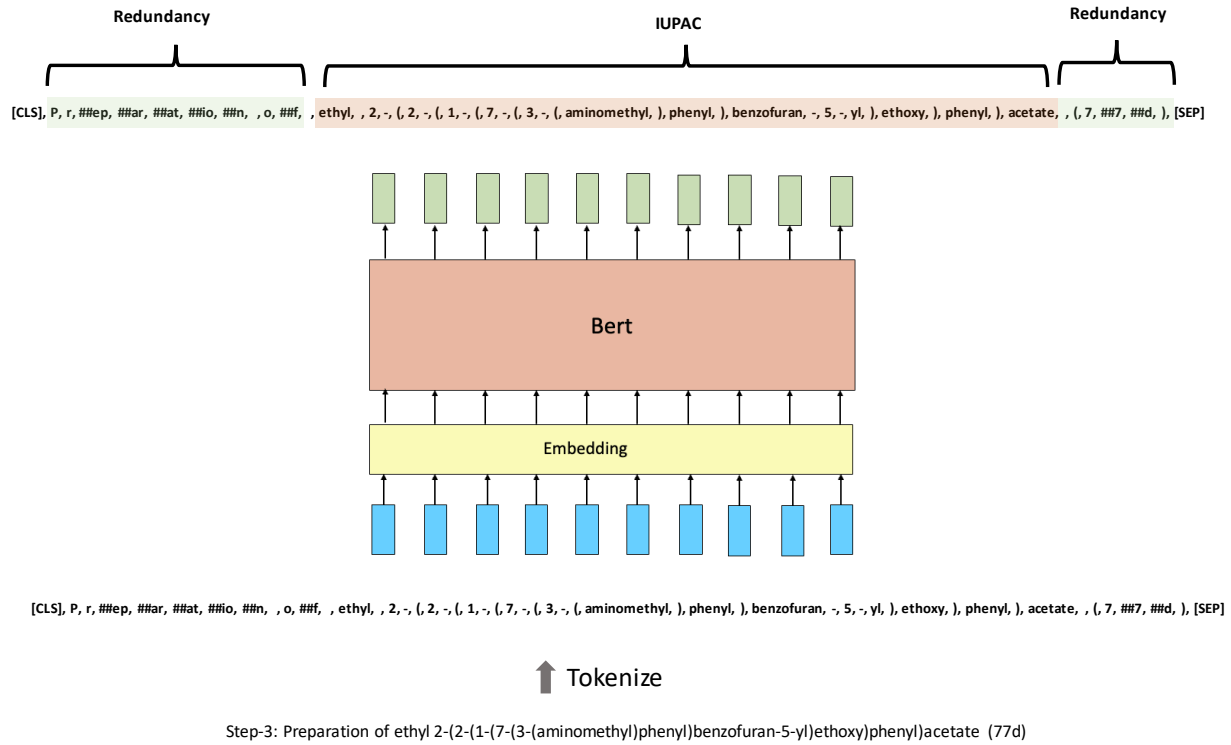
Typically, an OCR system is composed of two main components: text detection and text recognition. We employed the Differentiable Binarization (DB) technique (Liao, et al., 2020) for the text detection task and the Sequence-to-Sequence Visual Text Recognition (SVTR) algorithm (Du, et al., 2022) for the text recognition task. We fine-

tuned the pre-trained DB model using the manually-annotated dataset comprising 1,500 images sourced from patents and scientific literature.

We exploited the inherent structural dependencies within IUPAC nomenclature to enhance the accuracy of text recognition specifically for IUPAC strings. We generated a comprehensive dataset containing approximately 3.5 million images with IUPAC names embedded amidst random, redundant text, along with their ground truth labels. This dataset enables the model to discern the underlying structural relationships for a better character recognition performance. The IUPAC strings for this dataset were generated using 1.9 million SMILIES codes extracted from ChEMBL (Gaulton, et al., 2012). These SMILES codes were subsequently converted into IUPAC names using ChemDraw (Brown, 2014).

A detailed illustration of the complete OCR pipeline is provided in Supplementary Figure S1.

Named Entity Recognition



Supplementary Figure S2. Workflow of Chemical Named Entity Recognition. The figure delineates the complete process of chemical named entity recognition,

commencing with the pre-training of a BERT model on a corpus of 1.9 million IUPAC strings and culminating in a fine-tuned model designed for the task. The specialized tokenizer and output layer are integral elements of the fine-tuning stage.

Recent advancements in natural language processing (NLP) have spotlighted the efficacy of pre-training large-scale language models as a precursor to task-specific fine-tuning (Beltagy, et al., 2019). In alignment with this paradigm, we initially pre-trained a BERT (Devlin, et al., 2018) model on the in-house collected dataset comprising 1.9 million IUPAC strings. The pre-trained model was further fine-tuned by introducing an output layer specifically designed for chemical name entity recognition (NER).

The IUPAC nomenclature differs from the general text in both restricted lexicon and frequently recurred terms. Therefore, we eschewed the conventional WordPiece tokenizer found in the original BERT architecture. We developed a customized tokenizer tailored to the unique characteristics of the IUPAC nomenclature. The tokenizer was constructed on the principle of balancing a compact dictionary size with the capability for comprehensive encoding and decoding of IUPAC strings. This specialized approach effectively amplifies the accuracy of our IUPAC NER model. The resulting dictionary encapsulates a total of 1,500 distinct tokens, as shown in the Supplementary Table S3.

A schematic representation detailing the entire named entity recognition workflow is depicted in Supplementary Figure S2.

Error Correction Mechanism

Recognizing the sensitivity of IUPAC name conversion to OCR accuracy, we employ a dual-strategy error correction mechanism to improve recognition results. These strategies consist of a rule-based approach and a model-based technique.

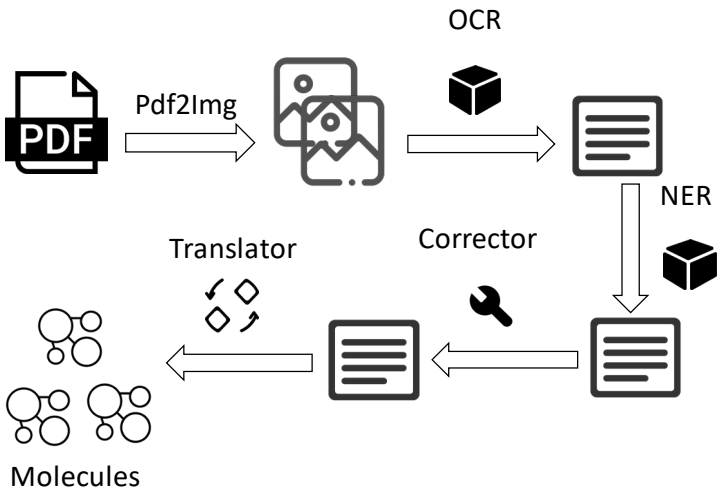
The rule-based system constructs a corpus of tuples representing commonly-occurred OCR errors alongside with their accurate counterparts, as depicted in Supplementary Table S1. The system scans the IUPAC strings via regular expressions to identify and replace erroneous substrings with their correct equivalents.

However, the stochastic nature of OCR errors makes it impractical to rely solely on predefined rules. The neural machine translation (NMT) technique (Wu, et al., 2016) has been successfully employed to fix OCR errors in various tasks (Mokhtar, et al., 2018) (Nastase and Hitschler, 2018). Inspired by the NMT strategy, this study incorporates a Transformer-based sequence-to-sequence model (Vaswani, et al., 2017) to rectify inaccuracies. While this approach significantly reduces the likelihood of version failures, it is essential to note the residual risk of incorrect or ambiguous transformation.

Chemical Structure Translator

The translation of the identified IUPAC names to chemical structures can be accomplished by several existing tools. STOUT is a deep-learning algorithm capable of bidirectional translation between SMILES strings and IUPAC names (Rajan, et al., 2021). Another tool molconvert is a command-line utility within the Marvin suite by ChemAxon (ChemAxon, 2016) to convert between various molecular file formats. After extensive comparison, we selected OPSIN (Lowe, et al., 2011), which provides superior accuracy in translating chemical names to structures while maintaining efficient inference speed.

Overall Architecture

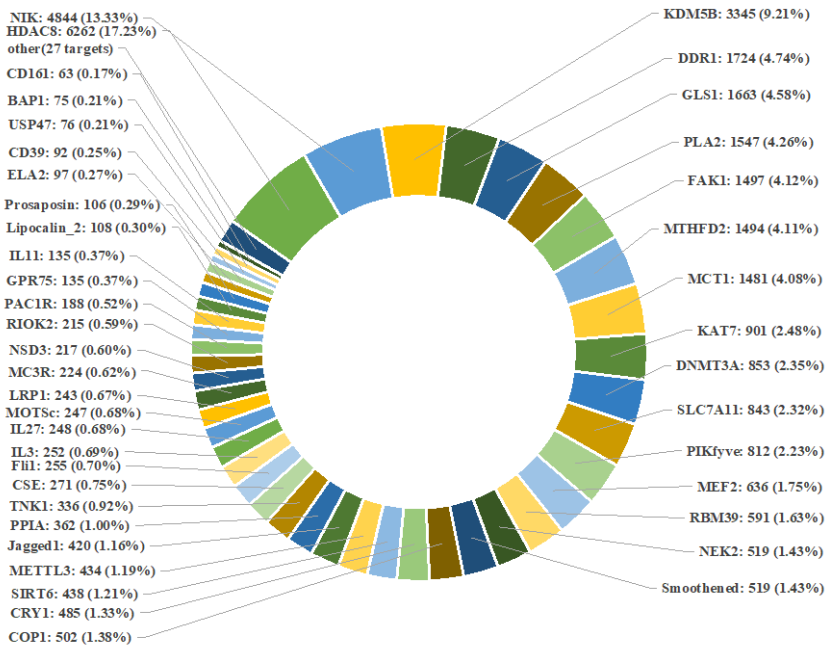


Supplementary Figure S3: Comprehensive Workflow of ChemDT. The figure

elucidates the end-to-end architecture of ChemDT, beginning with the conversion of scientific literature in PDF format to image data. It progresses through OCR, NER, error correction, and ultimately chemical structure translation using OPSIN.

The complete workflow of ChemDT is illustrated in Supplementary Figure S3. Given a scientific literature source in PDF format, we first convert it into a series of images. Each image undergoes text recognition via our OCR system and subsequently feeds into the NER model to identify IUPAC names across multiple text lines. These names then pass through the error correction mechanism for refinement before being converted into chemical structures (e.g., SMILES) using OPSIN. ChemDT is packed as a Docker image file and is freely available for non-commercial use.

Database Architecture



Supplementary Figure S4. Distribution of Therapeutic Targets and Corresponding Patent Literature. The figure showcases the variety and number of patents associated with each highlighted target.

Significant strides were made in the identification of novel therapeutic targets across a broad range of diseases in 2021. Milestone discoveries include anti-cancer targets like METTL3 (Yankova, et al., 2021), CD161 (Mathewson, et al., 2021), NSD3 (Yuan, et al., 2021), DDR1 (Sun, et al., 2021), ELA2 (Cui, et al., 2021), COP1 (Wang, et al., 2021), among others. Advances were also noted in other fields such as obesity, with newly identified targets like IL-27 (Wang, et al., 2021) and GPR75 (Akbari, et al., 2021). 2021 also saw the progress in diseases such as Alzheimer's disease, cardiovascular disorders, diabetes, and more. We conducted a survey of the most cited targets released in 2021. A curated list of 76 targets (Supplementary Table S4), corresponding to 942 patents and over 36,000 unique molecules. The distribution of these targets, along with their patent counts, is presented in Supplementary Figure S4.

We employed ChemDT to analyze all relevant patent documents and constructed the ChemDTD database. We first searched for the relevant target names in Google Patents, and then downloaded the corresponding PDF files. The architecture of ChemDTD is outlined in Supplementary Table S2. The current version of ChemDTD focuses on the high-impact targets in 2021, and is designed for scalability, allowing users to expand it via the ChemDT API.

Supplementary Table S1

Error-Truth Tuple Set for OCR Correction. The table lists examples of incorrectly recognized (OCRed) IUPAC substrings and their correct (Truth) counterparts.

OCRed	Truth	OCRed	Truth
methy!	methyl	aimethyl	dimethyl
pynido	pyrido	butyi	butyl
pyeazin	pyrazin	teIt	tert
earbamoyl	carbamoyl	lH	1H
telyl	tolyl	pyram	pyran
methanene	methanone	cthyl	ethyl
sulfony	sulfonyl	henzyl	benzyl

Supplementary Table S2

Schema of ChemDTD Database Fields. This table describes the various data fields available in the ChemDTD database, ranging from target names to molecular properties.

Field	Description
target	target name
patent	patent id
page	the page number of the recognized IUPAC
IUPAC	the recognized IUPAC name
smiles	SMILES of the recognized IUPAC name
MolWt	The average molecular weight of the molecule
NumHAcceptors	Number of Hydrogen Bond Acceptors
NumHDonors	Number of Hydrogen Bond Donors
MolLogP	Wildman-Crippen LogP value
NumRotatableBonds	Number of Rotatable Bonds

Supplementary Table S3

The dictionary obtained in the ChemDT NER module encapsulates a total of 1,500 distinct tokens.

Tokens
[PAD]
[UNK]
[CLS]
[SEP]
[MASK]
&
'
(
)
+
,
-
.
0
1
2
3
4
5
6
7
8
9
:
;
[
]
a
b
c
d
e

f
g
h
i
j
k
l
m
n
o
p
q
r
s
t
u
v
w
x
y
z
##i
##p
##h
##e
##n
##o
##x
##y
##s
##r
##l
##b
##u
##t
##a
##c
##f
##m
##d
##z
##v

##g
##k
##4
##q
##2
##9
##0
##8
##6
##1
##3
##5
##7
##j
##w
##yl
##th
##eth
##en
##ro
##in
##id
##ethyl
##an
methyl
##hen
yl
##ox
##henyl
##am
##rop
##ino
di
##az
##ol
##yr
##lo
##oxy
ox
##idin
eth
##methyl

pyr
##ar
##yd
##ide
am
##arb
phenyl
##azol
##amide
##ydro
ac
##er
##on
##anyl
##enz
##et
amino
prop
ethyl
##lu
##luo
carb
##ethoxy
##ic
##yc
benz
##yclo
##hlo
##phenyl
##ut
##ri
##ene
##ine
acid
methoxy
tri
chlo
##ul
##ulf
oxo
dimethyl
##ip

##prop
##amino
##iper
hydro
fluo
##hi
thi
##xy
##ano
but
##hydro
carbox
##im
cyclo
##at
piper
##idine
hydroxy
##ex
##rophenyl
##idene
##ent
##onyl
##ate
acet
##rol
##rom
##propyl
##etr
2s
##fluo
##ph
oxid
propan
benzo
oxidanyl
##sulf
##carb
chloro
tetr
##ne
dihydro

pyrrol
1h
az
##ane
##ranyl
brom
##hex
sulf
##azin
##ter
##thi
es
ester
propyl
butyl
##fluoro
##oyl
##is
##one
##it
carboxamide
pyridin
pyrim
tetra
amine
en
2r
##ur
carboxyl
one
10
##ec
##pyr
##pent
bromo
fluoro
##benz
##il
is
##but
##phen
##idyl

##di
oxidanylidene
##ep
11
oxy
bis
##idazol
thiazol
pyrazol
##cyclo
pent
trifluoro
##anoyl
##or
##eto
##amine
iso
##nd
ethoxy
thio
pyrimidin
im
meth
##uino
ter
tert
cyclohex
##carbox
##methoxy
ethan
fluorophenyl
hex
##hyd
##chlo
ind
##amoyl
##al
keto
12
acetamide
3r
13

##uinol
methoxyphenyl
nit
phen
3s
methan
trifluoromethyl
##os
##enyl
##uran
##thio
chlorophenyl
piperidin
##anoic
##ido
fluoranyl
benzamide
diox
##ilino
##dec
pyrrolidin
triazol
carboxylic
cyclopropyl
1s
##um
##icyclo
14
methylphenyl
tetrahydro
pyridyl
##ropyr
##aph
chloranyl
carbonyl
benzyl
##ept
##orph
##aphth
##orphol
1r
##sulfonyl

##ct
indol
##onamide
##piper
oxa
pyridine
an
sulfanyl
##yano
##hydropyr
##butyl
cyclopent
4s
imidazol
17
##diazol
carboxylate
15
##itri
##itril
##itrile
4r
cyano
##ac
nitro
enyl
oct
phenoxy
5r
morphol
dichlo
benzene
naphth
ylmethyl
##azine
piperazin
sulfonyl
16
methylamino
##ride
##furan
acetyl

azanyl
cyclohexyl
butan
##propan
dimethoxy
##od
quinol
##as
##imidazol
##carboxamide
carbam
##sulfanyl
methanone
trimethyl
dioxo
##inyl
5s
##idino
hept
##anoate
diaz
hexa
##osph
thiophen
2h
##iro
##piro
##ind
piperidyl
methylidene
##onitrile
piperidine
propanamide
pyrrolidine
pyridinyl
##phenoxy
##anilino
##benzo
##carboxyl
ol
oxoethyl
6r

dione
dimethylamino
hydroxymethyl
carbamoyl
acetic
pyrimidine
##eno
2
##deca
##indol
##thiazol
##carbonyl
##uin
##imid
oxazol
oxadiazol
ph
difluoro
##cyclohex
isopropyl
piperidinyl
##quinol
##chloride
##sulfonamide
anilino
##carboxylic
oxomethyl
1
benzodi
thia
##azo
##acyclo
##thiophen
##pyrrol
bromanyl
##acet
##ea
18
benzoic
propi
tetrahydropyr
tol

##ium
##oxazol
##benzyl
ethanamide
##pyrazol
phosph
##bicyclo
##rolo
phenylmethyl
naphthal
tolyl
benzimidazol
cyclopentyl
dihydroxy
methylene
iod
ethanone
quin
dimethylphenyl
##olan
hydrochloride
furan
19
6s
##azino
difluo
azido
diethyl
##onic
##uan
##ricyclo
dec
benzodiox
propanoyl
3a
dichlorophenyl
ium
diamine
##abicyclo
thienyl
ethylamino
dimethoxyphenyl

##carbamoyl
guan
##nyl
ur
methylsulfonyl
##amido
bromophenyl
##atricyclo
propanoic
pyrrolidinyl
nitrophenyl
octa
4h
##cyclohexyl
##hrom
diphenyl
##amethyl
benzothiazol
##pyridin
3h
urea
dien
non
propoxy
dichloro
tetrahydropyran
##butan
methylsulfanyl
20
guanidine
ylidene
##piperidin
imid
penta
imidazo
##uryl
21
quinazol
##iod
pyrid
pyrrolo
hyd

##hept
do
yloxy
pyrazolo
##quinolin
##pentyl
piperazinyl
pentan
methylpropyl
morpholin
carbamate
##osa
hydr
hydroxyphenyl
quinolin
iodo
tris
pyrazole
5h
##aspiro
benzenesulfonamide
benzoate
thiadiazol
difluorophenyl
##anth
##rophen
##ilyl
##piperazin
##iline
##propoxy
carbamic
piperazine
pyrazolyl
thiophene
ethane
dihydropyr
##iodide
hydroxyethyl
hydroiodide
##azep
azabicyclo
##imidin

##pyrrolidin
spiro
hexahydro
eno
methoxyethyl
##propane
benzofuran
pyrimidinyl
##rophenoxy
butoxy
oxan
aminomethyl
##keto
pyrazin
methylthio
acetate
ethanamine
fu
##ryl
ethanoyl
thiazole
furyl
##iso
4a
chrom
carbonitrile
##ren
##imino
ene
thiazolidin
propionamide
thieno
butanamide
tetrahydrofuran
acetamido
phenylethyl
22
quinoline
pentyl
indole
triazolo
phenylmethoxy

diketo
tetramethyl
6a
methylpropan
imino
methanamine
##oryl
##lyl
butanoyl
ylamino
ethenyl
benzodioxol
tetrazol
benzoyl
##de
pyrrole
quinazolin
benzox
aniline
##ot
##imidoyl
3
23
thiazolyl
cyclopenta
##benzene
naphthalen
##pyrimidin
naphthyl
hepta
##hydr
##benzamide
isoindol
##benzoyl
morpholino
pyridazin
triazin
##romethoxy
butyr
cyclohexane
oxolan
methanol

7a
azet
carboxy
cyclopropane
cyclohexa
isoxazol
##onium
sulfonamide
propanoate
cyclobutyl
ynyl
oxopropyl
8a
4
oxol
##carbonylamino
##orm
methylideneamino
##hydroxy
7r
propionic
sulfamoyl
ylphenyl
vinyl
trifluo
sulfanylidene
##ethan
##anium
azep
24
piperidino
butanoic
2e
morpholinyl
carbonylamino
form
piperazino
ethanol
##tetr
##sulfamoyl
##dehyd
##aldehyd

propane
##aldehyde
al
##aphthal
##phosph
carbazol
allyl
##morphol
trihydroxy
ylethyl
methylpyrazol
##acetyl
furanyl
##atetr
##ndec
benzothiophen
phenanth
##uinone
methoxymethyl
ethoxyphenyl
9s
oxanyl
##oxo
methane
methylol
dienyl
bicyclo
pyrrolidino
##atetracyclo
##oct
##naphthal
isoquinolin
##bd
pur
pyrido
silyl
enamide
7s
##bda
##ambda
##azinyl
oxymethyl

pentanoyl
##es
##quinoline
enoyl
n2
enoate
quinone
##pentan
nona
2z
phenol
##ca
isobutyl
##carbonitrile
##imidamide
benzoxazol
indazol
thiophenyl
undec
5z
##io
##hexyl
ethanoate
azetidin
##benzofuran
phosphoryl
dodeca
ethylphenyl
25
##cyclopropyl
##ynyl
##azepin
##acetamide
octan
diazep
##fu
hexan
##esyl
##monium
acryl
hexyl
##penta

benzodioxin
sulfonylamino
diethylamino
methylbut
##pentacyclo
benzonitrile
##enitrile
n1
##hydraz
tride
oxolanyl
ethylidene
oxopropan
trien
26
propylamino
triazole
##oxal
4e
##propanoyl
naphthalenyl
phenylphenyl
n4
ylsulfonyl
thiazolo
##naphthalen
8r
##urea
enoxy
8s
##enoic
##onate
triaz
5e
##oxymethyl
dioxa
quinolyl
3as
##tri
octahydro
methylphenoxy
cyclohept

##benzoic
chromen
diazaspiro
trimethoxy
azatricyclo
##pyran
##ant
##dio
diol
benzothi
imidazolidin
heptan
##brom
decan
cyclopentane
purin
##imidazo
10s
##hydrazide
##silyl
phenanthren
27
diamino
ethynyl
diene
morpholine
pentanoic
##carbam
##ric
ad
ethanoic
cyclopropylmethyl
14s
13s
butylphenyl
hydrazinyl
##quin
##chrom
azaspiro
##cosa
10r
oxazole

inden
##amant
ammonium
28
trifluoromethoxy
methylpiperidin
##inamide
pyrazine
##benzoate
##cyclopenta
##morpholin
enoic
cyanophenyl
3ar
##kis
methylprop
4z
hydrox
adamant
##lambda
thioxo
methylpiper
mesyl
##propanamide
##cyclopentyl
9r
##ipec
##bromo
carbo
iodanyl
##isoindol
yn
##carbamoylamino
naphthalene
butyric
imidazolyl
butane
##apentacyclo
dicarboxamide
cyclohexan
##aler
##thioyl

##oxyphenyl
##ycarb
methylcyclohexyl
deca
##decyl
tric
##ono
benzimidazole
3e
quinoxal
carbamoylamino
naphthyr
##icosa
methylpiperazin
difluoromethyl
methylpyridin
methylanilino
methyleneamino
dibromo
##amidine
##tetra
benzoxazin
##sulfonic
valer
difluoromethoxy
##butane
formyl
methylbutyl
##spiro
butanoate
##roprop
##ont
aminoethyl
##furo
##imidine
##ir
##ideneamino
diazatricyclo
benzopyran
propen
dihydroindol
imidazole

pyridylmethyl
butyramide
isopropoxy
##ipecot
7h
methylimidazol
##acetic
9a
##methan
tetradeca
methoxyphenoxy
##ycarbonyl
propanamine
##rene
##acont
diazepan
aminophenyl
pyridinecarboxamide
13r
isoquinoline
chloromethyl
oxycarbonyl
ylmethoxy
trideca
oxobutyl
dimethylpyrazol
methylbutan
##carboxylate
##sulfinyl
methoxyethoxy
##thal
##benzothiophen
dihydrochloride
##anoyloxy
##benzimidazol
thiourea
fluoren
fluorophenoxy
hydroxypropyl
3z
hexahydropyr
adamantyl

##triazol
azanium
decane
tetraen
chloride
phenylpropyl
9h
##pyrrolidine
oxadiazole
carbamimidoyl
##br
30
29
anis
chlorophenoxy
##ycyclo
##isoxazol
diethoxy
##idinone
benzenesulfonyl
butylamino
##inic
fluorobenzyl
##phosphoryl
azepan
5
##thieno
quinolinyl
methoxypropyl
nic
6ar
##ilane
trifluoroethyl
oxazolidin
dimethylphenoxy
##gen
nonan
furo
##uter
##uterio
##pyridine
tetraz

##cyclohexa
##oxan
chlorobenzyl
##othioyl
pyridazine
##thiolan
##decan
methylpent
trimethoxyphenyl
##dioic
triazine
##quinazol
##anoylamino
##ero
anth
nicot
piperidinecarboxamide
methanesulfonamide
6h
benzylidene
thiazolidine
benzyloxy
benzothiophene
17r
##furyl
oxobutan
##rac
carbaldehyde
pyran
11s
6as
##butoxy
##acosa
##sulfonate
trimethylsilyl
oxane
1e
propanoylamino
oxoprop
##cap
##to
heptane

pyrrolyl
fur
thione
pentanamide
##ercap
sulfo
##odi
7ar
##ercapto
##cyclobutyl
##piperazino
dimethylpropyl
##ethenyl
carbazole
azo
hexadeca
phthal
15
cyclopropylamino
phenylprop
##ylamino
mercapto
ethanoylamino
hydrazine
4as
ylsulfanyl
hexanoic
naphthyridin
n6
##quinazolin
triene
propanol
n3
phosphate
31
octadeca
diazabicyclo
##tert
12s
hydroxycyclo
ylpropyl
oxobut

oxido
cyclobutane
methylthiazol
##piperidine
##ropyridin
isoxazolyl
##cyano
valeric
5a
4ar
octane
pentaen
methylbenzyl
hydrazino
##aniline
##ethane
dithi
pentamethyl
indolyl
propanone
acetyloxy
##azepine
ditert
##onimidoyl
##imine
sec
##chromen
##azinan
methylbenzene
11r
methylcarbamoyl
dihydropyrazol
sulfanylmethyl
lambda
carboximidamide
trichlo
octadec
anisyl
triphenyl
benzenecarbonitrile
thiol
##piperidino

heptadeca
piperidinecarboxylic
33
##bor
##odium
dicarboxylic
pentane
32
benzodiazepin
tetrahydronaphthalen
propenoic
##ato
hydrogen
octyl
imidazolidine
13
isoxazole
cyclopropanecarboxamide
ethylsulfanyl
isoindole
ylmethylamino
silane
##iz
carbamothioyl
iodophenyl
##cos
dioxol
dicarboxylate
##imido
##octa
##bs
##carboximidamide
##diamide
oxolane
ethylsulfonyl
14r
17s
##hydrocyclopenta
##aconta
12r
trimethylphenyl
trione

##ipecotamide
tolylmethyl
ylcarbonyl
indolin
fluoroanilino
methylthiophen
propenamide
iumyl
##oxolan
tetral
tetralin
iminomethyl
quinoxalin
acrylamide
benzoxy
benzimidazolyl
quinazoline
10a
##furanyl
methoxycarbonyl
pyrimidinamine
bromomethyl
##tetrazol
##hexan
##do
##azepan
pyrrolidinecarboxamide
35
##pyrimidine
8
isoquinol
##umar
ethylthio
anthrac
fluorene
8as
oxycarbonylamino
azetidine
##carbo
dodec
methylpyrrolidin
pentadeca

co
tetrakis
oxopent
##hexacyclo
##onamido
8ar
butoxycarbonylamino
yloxyphenyl
9z
##thioamide
##carbazol
##dioate
##azet
##lambda6
dihydropyrrolo
oxabicyclo
##methylamino
enylidene
acrylic
methylcyclopropyl
guanidino
undecan
dioxolan
dimethylanilino
34
11
hexane
heptanyl
oxopentyl
indan
##icos
##etyl
ison
carboxamidine
tetraene
cycloheptyl
o1
7as
docosa
##epin
pyrrolidinecarboxylic
cyclopentan

phenoxyphenyl
36
pic
diaminomethyl
hexanoyl
thiolan
nonadeca
trifluoroacetic
methylimino
oxopentan
tricyclo
indoline
dioxan
##hexa
##ethylamino
heptyl
oxir
triazatricyclo
chromene
diiso
##pyrrolidino
##sulfonamido
15s
7z
carbonimidoyl
##propylamino
phenylpropan
oxyphenyl
methylaminomethyl
chloroethyl
37
methylpentan
acetonitrile
1z
tetrazolyl
hen
##azanium
diazenyl
methylbenzamide
##hydrazinyl
##urin
##acos

dihydrobenzofuran
##isoquinolin
methoxyanilino
15r
dibenzofuran
azetidinyl
##idazin
hexadec
pyrazinyl
ethylideneamino
##anide
triol
hexaen
isocyano
hydrazinylidene
##azinane
oxet
cyclohexylmethyl
oxyethyl
##benzenesulfonamide
##hepta
xanth
cyclopropylmethoxy
methylsulfinyl
phenetyl
cyclopropan
11a
pentanoate
##ethanamine
trifluoroethoxy
12
acetoxy
trichloro
methylfuran
phenylmethoxyphenyl
##carbamic
dimethylsulfamoyl
##oxa
triazolyl
dimethylcyclohexyl
##oni
cyclohepta

phenylpyrazol
picol
methylpyrimidin
39
din
##bo
methylpropoxy
16r
thiadiazole
pyridinamine
oxoethoxy
##allyl
methylsulfonylphenyl
propoxyphenyl
azidooxy
amyl
##piperazine
##ethi
nonane
##sulfonylamino
sodium
benzotriazol
icosa
##azetidin
indazole
oxamide
##yan
##hydrop
trifluoromethyloxy
cyclohexanecarboxylic
cyclohexanecarboxamide
dodecahydro
##octyl
9b
##ron
phenylsulfanyl
diazatetracyclo
benzothiazole
cyanomethyl
phosphono
pyridinecarboxylic
ylthio

##so
coumar
##ethio
6e
##orbor
undeca
furfuryl
pyridinylmethyl
oxazolyl
12a
dioxothiolan
cyclohexen
##cin
##acetate
16s
methylpiperazino
13z
ethoxycarbonyl
14
chroman
##guan
11z
##guanidine
carbohydrazide
##imidazolidin
##methanamine
pyridazinyl
##propanoic
quinazolinyl
diaminomethylideneamino
tetradec
tetrahydrobenzothiophen
##ropropyl
methylindol
imine
##purin
besyl
sulfonic
##heptyl
hexanamide
azepin
dihydroind

6z
##butanoyl
aza
dinit
nicotinamide
##eroxy
trienyl
##quinoxal
benzofuranyl
tetrafluoro
azapentacyclo
##tetrahydro
##el
phosphanyl
hene
12z
dichlorobenzyl
benzamido
40
##butanamide
hydroxyethoxy
9
16
cyanoethyl
##acetyloxy
isoindolin
propanenitrile
diazo
ethoxymethyl
thiazolidinyl
##heptan
benzoxazine
chloroanilino
10b
decahydro
methylsulfanylphenyl
benzazepin
##enoate
ap

Supplementary Table S4

The curated list of 76 disease targets with the most citations in 2021.

Target
GPR171
PUS7
TRAPPC4
CHMP7
ABCB10
CEACAM7
DNPH1
Neuraminidase_1_NEU1
VPS39
RSK4
ALYREF
HRH1
HULC
MOTSc
CNNM4
FGD5
MCAD
SerpinB13
Cop1
TGFbeta2_AND_TGFbeta3
BTN1A1
STC1
FAK1
Tim4
GPR75
KAT7
USP47
TNK1
ANLN
IL17RB
NSD3
PAC1R
SETDB1
microRNA21
MC3R
YTHDF2

MTHFD2
RIOK2
Jagged1
B4GALT1
CD93
IGF2BP3
BAP1
SAMHD1
Prosaposin
CD161
CRY1
DAXX
SOCS1
RBM39
ELA2
GLS1
METTL3
NEK2
MCT1
LRP1
MEF2
Fli1
CSE
SIRT6
SLC7A11
DNMT3A
KDM5B
CD39
Lipocalin_2
IL11
IL27
PIKfyve
IL3
HDAC8
NIK
SHP2
Smoothened
PPIA
DDR1
PLA2

References

Akbari, P.*, et al.* Sequencing of 640,000 exomes identifies GPR75 variants associated with protection from obesity. *Science* 2021;373(6550):eabf8683 %@ 0036-8075.

Beltagy, I., Lo, K. and Cohan, A. SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676* 2019.

Brown, T. ChemDraw. *The Science Teacher* 2014;81(2):67 %@ 0036-8555.

ChemAxon. ChemAxon–Software solutions and services for chemistry and biology. In.: ChemAxon; 2016.

Cui, C.*, et al.* Neutrophil elastase selectively kills cancer cells and attenuates tumorigenesis. *Cell* 2021;184(12):3163-3177. e3121 %@ 0092-8674.

Dalby, A.*, et al.* Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *Journal of chemical information and computer sciences* 1992;32(3):244-255 %@ 0095-2338.

Devlin, J.*, et al.* Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* 2018.

Du, Y.*, et al.* SVTR: Scene Text Recognition with a Single Visual Model. *arXiv preprint arXiv:2205.00159* 2022.

Du, Y.*, et al.* Pp-ocr: A practical ultra lightweight ocr system. *arXiv preprint arXiv:2009.09941* 2020.

Fleming, N. How artificial intelligence is changing drug discovery. *Nature* 2018;557(7706):S55-S55 %@ 0028-0836.

Gaulton, A.*, et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research* 2012;40(D1):D1100-D1107 %@ 1362-4962.

Ghasemi, F.*, et al.* Neural network and deep-learning algorithms used in QSAR studies: merits and drawbacks. *Drug discovery today* 2018;23(10):1784-1790 %@ 1878-5832.

Heller, S.R.*, et al.* InChI, the IUPAC international chemical identifier. *Journal of cheminformatics* 2015;7(1):1-34 %@ 1758-2946.

Holzinger, A., Haibe-Kains, B. and Jurisica, I. Why imaging data alone is not enough: AI-based integration of imaging, omics, and clinical data. *European Journal of Nuclear Medicine and Molecular Imaging* 2019;46(13):2722-2730 %@ 1619-7089.

Jin, W., Barzilay, R. and Jaakkola, T. Junction tree variational autoencoder for molecular graph generation. In.: PMLR; 2018. p. 2323-2332 %@ 2640-3498.

Lafferty, J., McCallum, A. and Pereira, F.C.N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

Li, J.*, et al.* A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering* 2020;34(1):50-70 %@ 1041-4347.

Liao, M.*, et al.* Real-time scene text detection with differentiable binarization. In.; 2020. p. 11474-11481 %@ 12374-13468.

Lowe, D.M.*, et al.* Chemical name to structure: OPSIN, an open source solution. In.: ACS Publications; 2011.

Mathewson, N.D.*, et al.* Inhibitory CD161 receptor identified in glioma-infiltrating T cells by single-cell analysis. *Cell* 2021;184(5):1281-1298. e1226 %@ 0092-8674.

Mokhtar, K., Bukhari, S.S. and Dengel, A. OCR Error Correction: State-of-the-Art vs an NMT-based Approach. In.: IEEE; 2018. p. 429-434 %@ 1538633469.

Nastase, V. and Hitschler, J. Correction of OCR word segmentation errors in articles from the ACL collection through neural machine translation methods. In.; 2018.

Oldenhof, M.*, et al.* ChemGrapher: optical graph recognition of chemical compounds by deep learning. *Journal of chemical information and modeling* 2020;60(10):4506-4517 %@ 1549-9596.

Patton, E.E., Zon, L.I. and Langenau, D.M. Zebrafish disease models in drug discovery: from preclinical modelling to clinical trials. *Nature Reviews Drug Discovery* 2021;20(8):611-628 %@ 1474-1784.

Rajan, K.*, et al.* A review of optical chemical structure recognition tools. *Journal of Cheminformatics* 2020;12(1):1-13 %@ 1758-2946.

Rajan, K., Zielesny, A. and Steinbeck, C. STOUT: SMILES to IUPAC names using neural machine translation. *Journal of Cheminformatics* 2021;13(1):1-14 %@ 1758-2946.

Rocktäschel, T., Weidlich, M. and Leser, U. ChemSpot: a hybrid system for chemical named entity recognition. *Bioinformatics* 2012;28(12):1633-1640 %@ 1460-2059.

Savage, N. Tapping into the drug discovery potential of AI. *Biopharma Deal* 2021.

Schenone, M.*, et al.* Target identification and mechanism of action in chemical biology and drug discovery. *Nature chemical biology* 2013;9(4):232-240 %@ 1552-4469.

Smith, R. An overview of the Tesseract OCR engine. In.: IEEE; 2007. p. 629-633 %@ 0769528228.

Sun, X.*, et al.* Tumour DDR1 promotes collagen fibre alignment to instigate immune exclusion. *Nature* 2021;599(7886):673-678 %@ 1476-4687.

Tan, X.*, et al.* Discovery of pyrazolo [3, 4-d] pyridazinone derivatives as selective

DDR1 inhibitors via deep learning based design, synthesis, and biological evaluation. *Journal of medicinal chemistry %@ 0022-2623* 2021.

Tsubaki, M., Tomii, K. and Sese, J. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics* 2019;35(2):309-318 %@ 1367-4803.

Usié, A.*, et al.* CheNER: chemical named entity recognizer. *Bioinformatics* 2014;30(7):1039-1040 %@ 1460-2059.

Vaswani, A.*, et al.* Attention is all you need. *Advances in neural information processing systems* 2017;30.

Wang, Q.*, et al.* IL-27 signalling promotes adipocyte thermogenesis and energy expenditure. *Nature* 2021;600(7888):314-318 %@ 1476-4687.

Wang, X.*, et al.* In vivo CRISPR screens identify the E3 ligase Cop1 as a modulator of macrophage infiltration and cancer immunotherapy target. *Cell* 2021;184(21):5357-5374. e5322 %@ 0092-8674.

Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* 1988;28(1):31-36 %@ 0095-2338.

Wu, Y.*, et al.* Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144* 2016.

Yankova, E.*, et al.* Small-molecule inhibition of METTL3 as a strategy against myeloid leukaemia. *Nature* 2021;593(7860):597-601 %@ 1476-4687.

Yoo, S., Kwon, O. and Lee, H. Image-to-Graph Transformers for Chemical Structure Recognition. In.: IEEE; 2022. p. 3393-3397 %@ 1665405406.

Yuan, G.*, et al.* Elevated NSD3 histone methylation activity drives squamous cell lung cancer. *Nature* 2021;590(7846):504-508 %@ 1476-4687.

Zhavoronkov, A.*, et al.* Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature biotechnology* 2019;37(9):1038-1040 %@ 1546-1696.