ChemPatentizer: Transforming Chemical Patents into Actionable Scientific Data

Riccardo Fusco MSc.
Institute for Molecular and Translational and Medicine (IMTM)
Czech Advanced Technology Research Institute (CATRIN)
Palacky University, Olomouc, Czechia

ChemPatentizer

Transforming Chemical Patents into Actionable Scientific Data

What is a Pharmaceutical Patent?

- A document that protect the intellectual propriety associated to the discovery of a new drug.
- *Strategically* a problematic tradeoff. They trade secrecy for exclusivity.
- All the interest to *hide* the information as much as possible (while keeping legal validity).

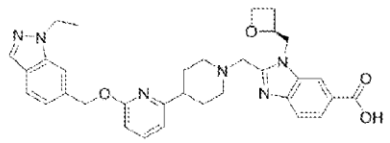
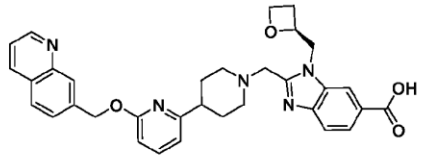
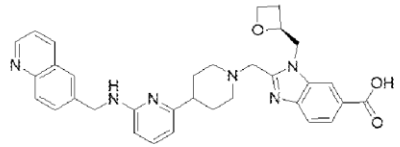
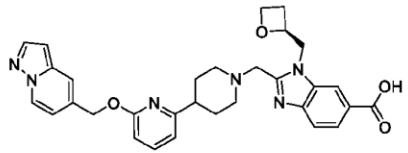
Why is so important for a Drug Hunter?

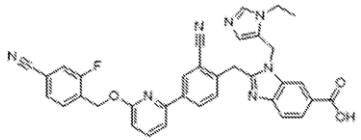
- Contains high quality structure-activity relationship
- An in-dept analysis gives opportunity for *industrial intelligence*

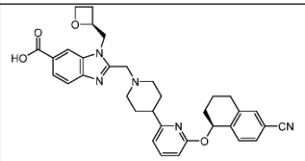
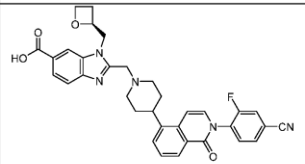
What ChemPatentizer aim to do?

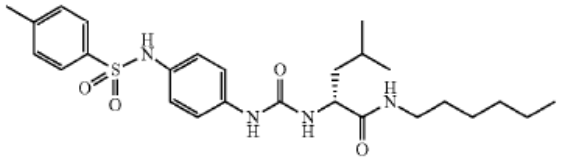
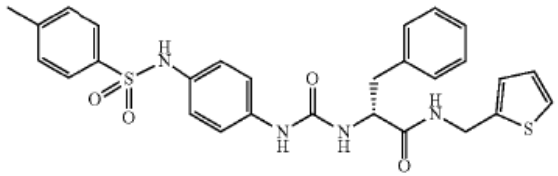
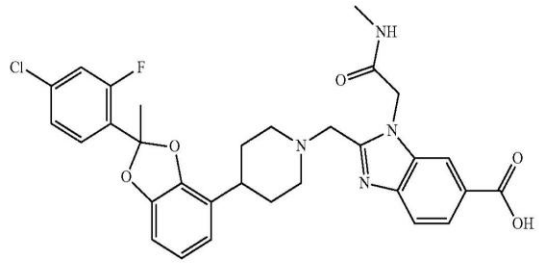
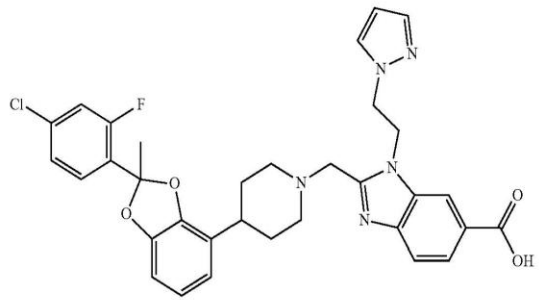
- Upload a user-defined list of PDF
- Extraction of the structure-activity data
- (Optional) Molecular matched pair (MMPA) for identifying the most impactful groups for the activity
- (Optional) *Deep*-QSAR modelling to predict activity of user data

Challenges – Structure Wise

21		83	
22		84	

522	 2-(2-cyano-4-(6-((4-cyano-2-fluorobenzyl)oxy)pyridin-2-yl)benzyl)-1-((1-ethyl-1H-imidazol-5-yl)methyl)-1H-benzo[d]imidazole-6-carboxylic acid: ES/MS m/z 612.2; ^{1}H NMR (400 MHz, Methanol-d4) δ 9.31 (s, 1H), 9.03 (s,

Cmpd No.	Structure	Name
146-P2		2-((4-(6-(((S)-6-cyano-1,2,3,4-tetrahydronaphthalen-1-yl)oxy)pyridin-2-yl)piperidin-1-yl)methyl)-1-(((S)-oxetan-2-yl)methyl)-1H-benzo[d]imidazole-6-carboxylic acid
147		(S)-2-((4-(2-(4-cyano-2-fluorophenyl)-1-oxo-1,2-dihydroisoquinolin-5-yl)piperidin-1-yl)methyl)-1-(oxetan-2-ylmethyl)-1H-benzo[d]imidazole-6-carboxylic acid

comp. No.	structure	MS (ESI) m/z (M + H)$^+$	NMR
11		503	1H NMR (400 MHz, DMSO-d6) δ 9.81 (s, 1H), 8.48 (s, 1H), 8.02 (t, J = 5.7 Hz, 1H), 7.59-7.51 (m, 2H), 7.31 (dd, J = 8.1, 0.9 Hz, 2H), 7.24-7.16 (m, 2H), 6.95-6.86 (m, 2H), 6.22 (d, J = 8.5 Hz, 1H), 4.22-4.10 (m, 1H), 3.13-2.92 (m, 2H), 2.32 (s, 3H), 1.65-1.50 (m, 1H), 1.48-1.31 (m, 4H), 1.24 (q, J = 4.6 Hz, 6H), 0.91-0.81 (m, 9H).
12		549	not measured
52			2-({4-[2-(4-chloro-2-fluorophenyl)-2-methyl-1,3-benzodioxol-4-yl]piperidin-1-yl}methyl)-1-[2-(methylamino)-2-oxoethyl]-1H-benzimidazole-6-carboxylic acid
53			2-({4-[2-(4-chloro-2-fluorophenyl)-2-methyl-1,3-benzodioxol-4-yl]piperidin-1-yl}methyl)-1-[2-(1H-pyrazol-1-yl)ethyl]-1H-benzimidazole-6-carboxylic acid

IMTM

Challenges – Table Wise

表一：本发明化合物对人 GLP-1 R 的 EC_{50}

实施例	EC_{50} (nM)
1	0.48
2	1.94
3	5.56
4	1.57
5	2.59
7	2.86
11	2.01
14	3.05
15	0.99
16	1.88
18	8.00
19	6.07
20	6.92
21	1.70
23	7.44
24	4.49
27	2.76
29	6.69

Example	**GLP-1R Cell Assay 1 EC_{50} (nM)**
36	36.4418
37	>500
38	>500
39	420.1650
40	3.2834
41	>500
42	3.8409
43	>500
44	82.1240
45	21.6840
46	>500
47	>500
48	>500
49	>500

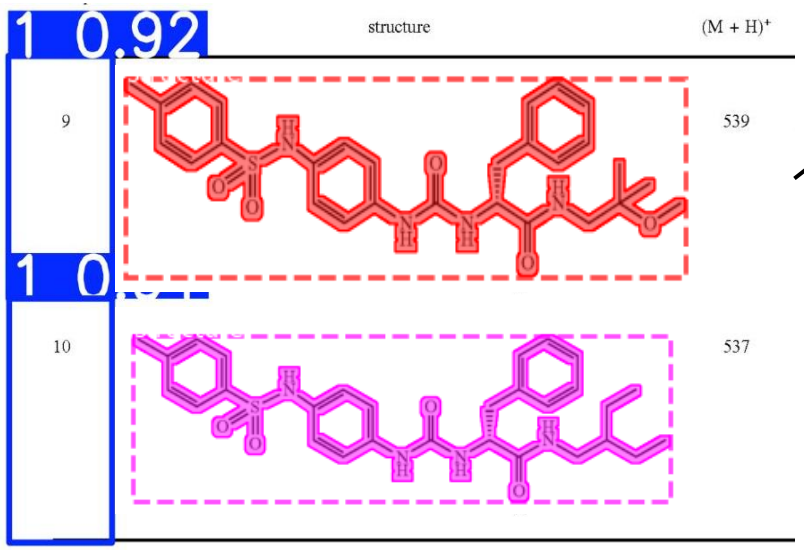
TABLE 11-continued

Biological activity for Examples 1-80.

Example Number	Assay 1 EC_{50} (nM)	Assay 1 Emax (%)	Assay 1 Number	Assay 2 EC_{50} (nM)	Assay 2 Emax (%)	Assay 2 Number
14	14	77	3	540	120	3
15	0.96	81	4	21	91	3
16	0.99	87	3	18	130	4
17	6.0	86	3	150	91	3
18	1.8	95	3	59	100	3
19	5.3	90	8	42	94	6
20	0.34	80	5	6.1	91	4
21	14	73	3	370	91	3
22	2.8	85	5	23	82	3
23	41	89	4	450	94	4
24	2.0	76	3	33	89	3
25	6.3	80	4	73	91	3
26	5.1	86	4	38	86	3
27	0.84	86	4	16	85	5
28	6.8	94	3	150	98	3
29	1800	65	4			
30	140	79	3	2800	110	4
31	300	65	3			
32	>20000		1			
33	3.1	93	3	84	110	3
34	1900	92	3			
35	330	85	3	13000	100	3
36	0.48	83	3	15	90	5
37	9.3	88	3	190	92	3
38	1.1	75	6	35	91	7
39	1.6	90	3	29	95	3
40	150	77	3	2000	100	3
41	7.6	84	4	130	96	3
42	3400	89	3			

IMTM

Solution



DECIMER [1] Image Segmentation

- Find and segment the chemical structures

DECIMER [1] Image Transformer

- Convert from image to SMILES

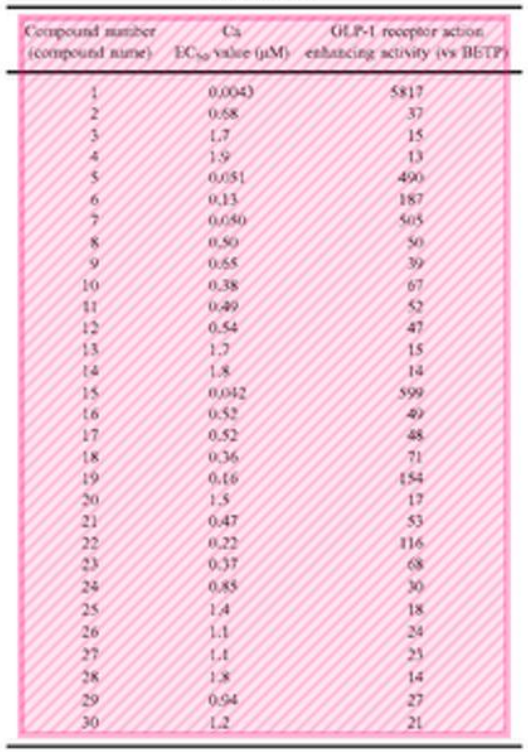


TABLE 4-1

Compound number (compound name)	Ca_v EC_{50} value (μM)	GLP-1 receptor action enhancing activity (vs BETP)
1	0.0043	5817
2	0.58	37
3	1.7	15
4	1.9	13
5	0.051	490
6	0.13	187
7	0.050	505
8	0.50	50
9	0.55	39
10	0.38	67
11	0.49	52
12	0.54	47
13	1.7	15
14	1.8	14
15	0.042	599
16	0.52	49
17	0.52	48
18	0.36	71
19	0.16	154
20	1.5	17
21	0.47	53
22	0.22	116
23	0.37	68
24	0.85	30
25	1.4	18
26	1.1	24
27	1.1	25
28	1.8	14
29	0.94	27
30	1.2	21

Table Transformer (TATR) [2]

- Extract the table from the wild

Mistral OCR [3]

- Convert table (image) to Markdown

Confidential

- Find the ID
- Trainable by the User

OCR Recognition (Whatever)

- Convert the ID to actual text

[1] Rajan, K., Brinkhaus, H. O., Agea, M. I., Zielesny, A., & Steinbeck, C. (2023). DECIMER.ai: an open platform for automated optical chemical structure identification, segmentation and recognition in scientific publications. Nature Communications, 14, 5045. https://doi.org/10.1038/s41467-023-40782-0

[2] Smock, B., Pesala, R., & Abraham, R. (2022, June). PubTables-1M: Towards comprehensive table extraction from unstructured documents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 4634–4642).

[3] Mistral AI Team. (2025, March 6). Mistral OCR: Introducing the world's best document understanding API. *Mistral AI*. https://mistral.ai/news/mistral-ocr

The Trick

```
############################################################
[Patent_Descriptions.Ajinomoto_20191227]
Assay = "Ca_GLP1_Assays"
Mode = "PDF"

Freedom = 180
ID_Side = ["left"]          # They can be "left", "right" "top"
Page_Ranges = [
    { start = 50, end = 71, tag = "1" }
]
[[Patent_Descriptions.Ajinomoto_20191227.Activity_Descriptors]]
Pages = [73]
Assay Pages = [72, 74]
Automatic_Extraction = false

Activity_Columns = [
    "Compound number (compound name)",
    "Ca EC50 value (μM)",
    "GLP-1 receptor action enhancing activity (vs BETP)",
]
[redacted]
Unit = "μM"
[redacted]
############################################################
```

```
"Astrazeneca_20221215": {
    Assay : {
        "CHOK1_GLP1R_cAMP_Assay": "QSAR_METHOD_1"
    }
},
"CelgeneInternational_20151210": {
    Assay : {
        "GLP1R_PAM_Shift_cAMP_Assay": "QSAR_METHOD_1",
        "GLP1R_CRE_bla_CHO_HI_cAMP_Assay": "QSAR_METHOD_1"
    }
},
"CelgeneInternational_20180426": {
    Assay : {
        "EC2GLP-1(9-36)PAMcAMPAssay": "QSAR_METHOD_1"
    }
},
"ChugaiSeiyakuKabushikiKaisha_20170926": {
    Assay : {
        "InVitroCAMPActivationAssay": "QSAR_METHOD_1"
    }
},
"Gasherbrum_20210205": {
    Assay : {
        "GLP1Receptor_cAMP_Stimulation_Assay": "QSAR_METHOD_1",
        "GLP1Receptor_EFC_Assay": "QSAR_METHOD_1",
        "GLP1Receptor_HEK293_CRELUC_Assay": "QSAR_METHOD_1",
        "Rat_Pharmacokinetics_Study": "QSAR_METHOD_4"
    }
},
```

IMTM

Results

Document	Segment	ID	Score	Assay Type	SMILES
Ajinomoto_20191227_extracted	page_1_segment_0	1	96.0	Ca_GLP1_Assays	CCCCC(C(
Ajinomoto_20191227_extracted	page_2_segment_0	2	100.0	Ca_GLP1_Assays	CCCCCCN
Ajinomoto_20191227_extracted	page_2_segment_1	3	100.0	Ca_GLP1_Assays	Cc1ccc(S(=
Ajinomoto_20191227_extracted	page_2_segment_2	4	96.0	Ca_GLP1_Assays	Cc1ccc(S(=
Ajinomoto_20191227_extracted	page_2_segment_3	5	100.0	Ca_GLP1_Assays	CCCCCCC

MMPA, DeepQSAR or whatever it's a CSV

Validation

As always, it's a problem

- Manual validation
  Unfeasible
- SciFinder and Reaxys API
  Needs actual collaboration because they don't release them easily

Call to Action

Thanks for the Attention