technische universität dortmund

fakultät für chemie und chemische biologie

Datenkompetenz in der Chemie und Chemischen Biologie

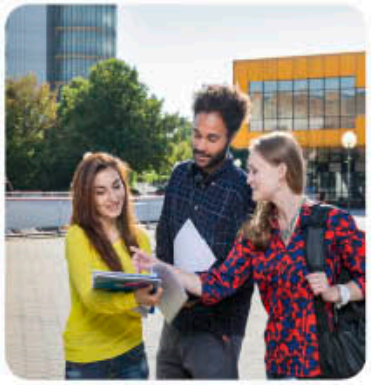
Prof. Dr. Paul Czodrowski

www.czodrowskilab.org

twitter.com/czodrowskipaul

www.czodrowskilab.org/music

github.com/czodrowskilab



WiSe **Statistik in den Lebenswissenschaften**

SoSe **Einführung Data Science im Bereich Chemie und Chemische Biologie**

WiSe **[*In Planung*] Statistisches Lernen**

Alle Veranstaltung: 2 V + 1 Ü

WiSe

Statistik in den Lebenswissenschaften

Thema	Inhalte (Thematische Unterteilung)	Schlüsselbegriffe
Was ist Statistik?	Einführung Statistik, Organisatorisches, Merkmale, Grundlagen der Mathematik	Python, Logarithmus, Wurzeln, Funktionen
Deskriptive Statistik	Beschreibung von Daten mittels der Deskriptiven Statistik	Gini-Koeffizient, Lorenzkurve, Häufigkeiten, Median, MAD (mean absolute deviation)
Deskriptive Statistik	Beschreibung von Daten mittels der Deskriptiven Statistik	Kendall's tau, Standardabweichung, Varianz, Mittelwerte, Histogramm
Korrelation und Regression	Rangkorrelation, Regression	Spearman, Scheinkorrelation, Verdeckte Korrelation
Klassierte Daten	Umgang und Interpretation von kategorialen Daten	Kontingenztafeln, chi2, odds ratio
Sensitivität und Spezifität	Sensitivität und Spezifizität, External quality assessment, hERG kapa	Sensitivität und Spezifität, kappa
Fallbeispiele	Fallbeispiel SARS-CoV2, hERG	kappa, Hypothesen-Test logP bei hERG
Fallbeispiele	hERG, Biochemie-Praktikums-Versuch	MatchedMolecularPairs & Significance, Michaelis-Menten & Error

Beispiel: Lineare Regression

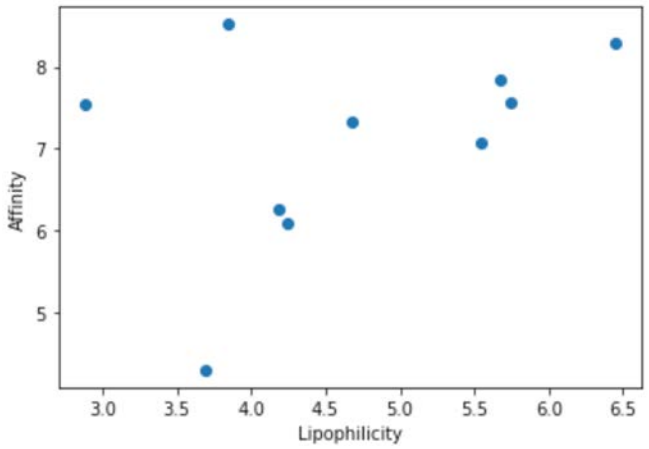
Zitat

If your experiment needs a statistician, you need a better experiment.

Ernest Rutherford

Plausibilitätsabschätzungen

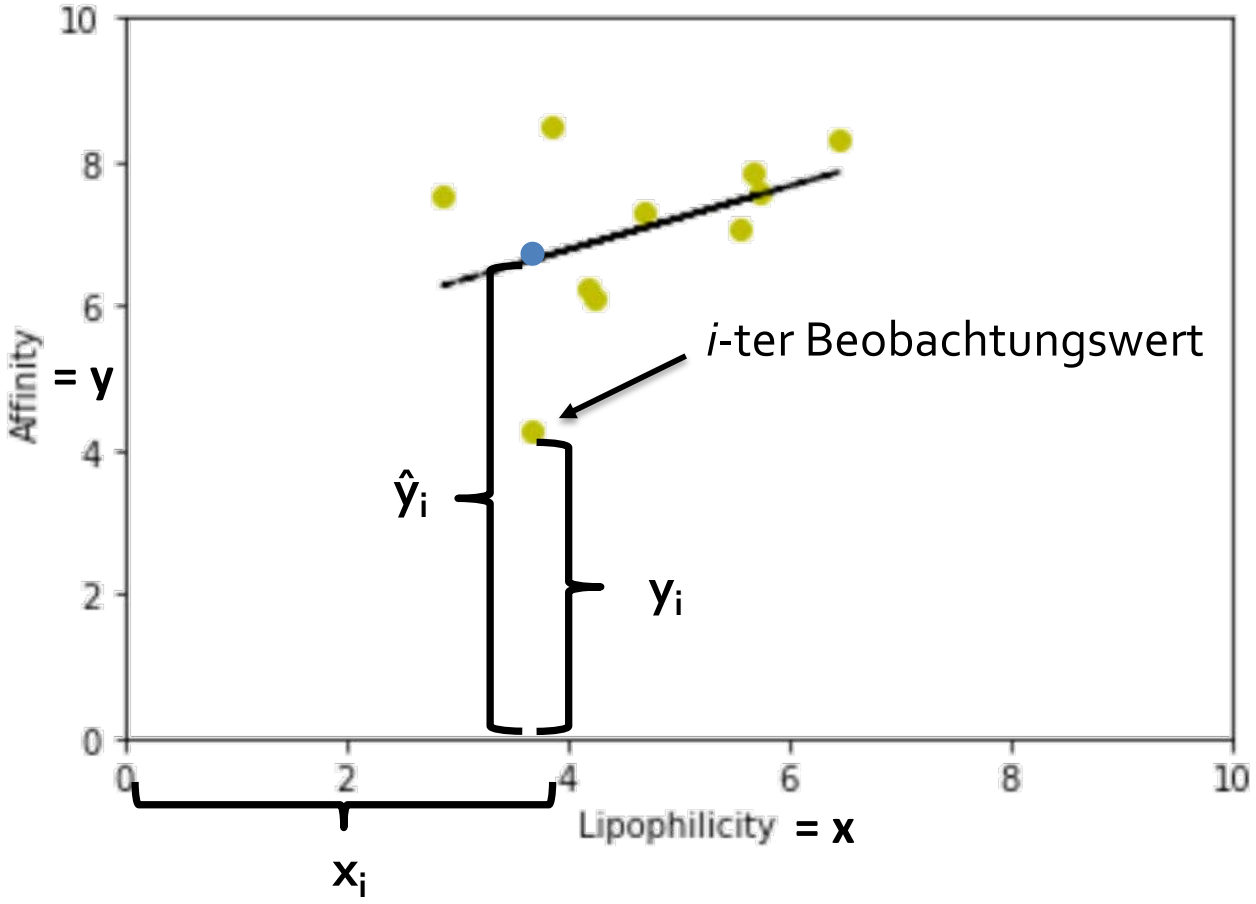
	Molekül1	Molekül2	Molekül3	Molekül4	Molekül5	Molekül6	Molekül7	Molekül8	Molekül9	Molekül10
Lipophilie	3,84	5,67	5,55	2,88	4,24	3,69	6,45	4,18	4,68	5,74
Affinität	8,52	7,85	7,08	7,54	6,09	4,29	8,3	6,26	7,32	7,57



- Sie haben 10 Moleküle im Labor synthetisiert und die Affinität auf dem *Target* p38 bestimmt
- Für diese 10 Moleküle haben Sie ebenso die Lipophilie (der sogenannte logP-Wert) bestimmt
- Mittels Bravais-Pearsion erhalten Sie eine Korrelation von 0,40

Lässt sich nun für ein neues Molekül (bspw. mit einem logP=5) die Affinität auf p38 abschätzen?
Anders ausgedrückt: wie hängen logP und pIC50 linear miteinander zusammen?
Aus Plausibilitätsbetrachtungen können Sie auf einen Zusammenhang von pIC50 und logP schließen, da die Bindetasche von p38 einen stark lipophilen Charakter hat

Lineare Regression

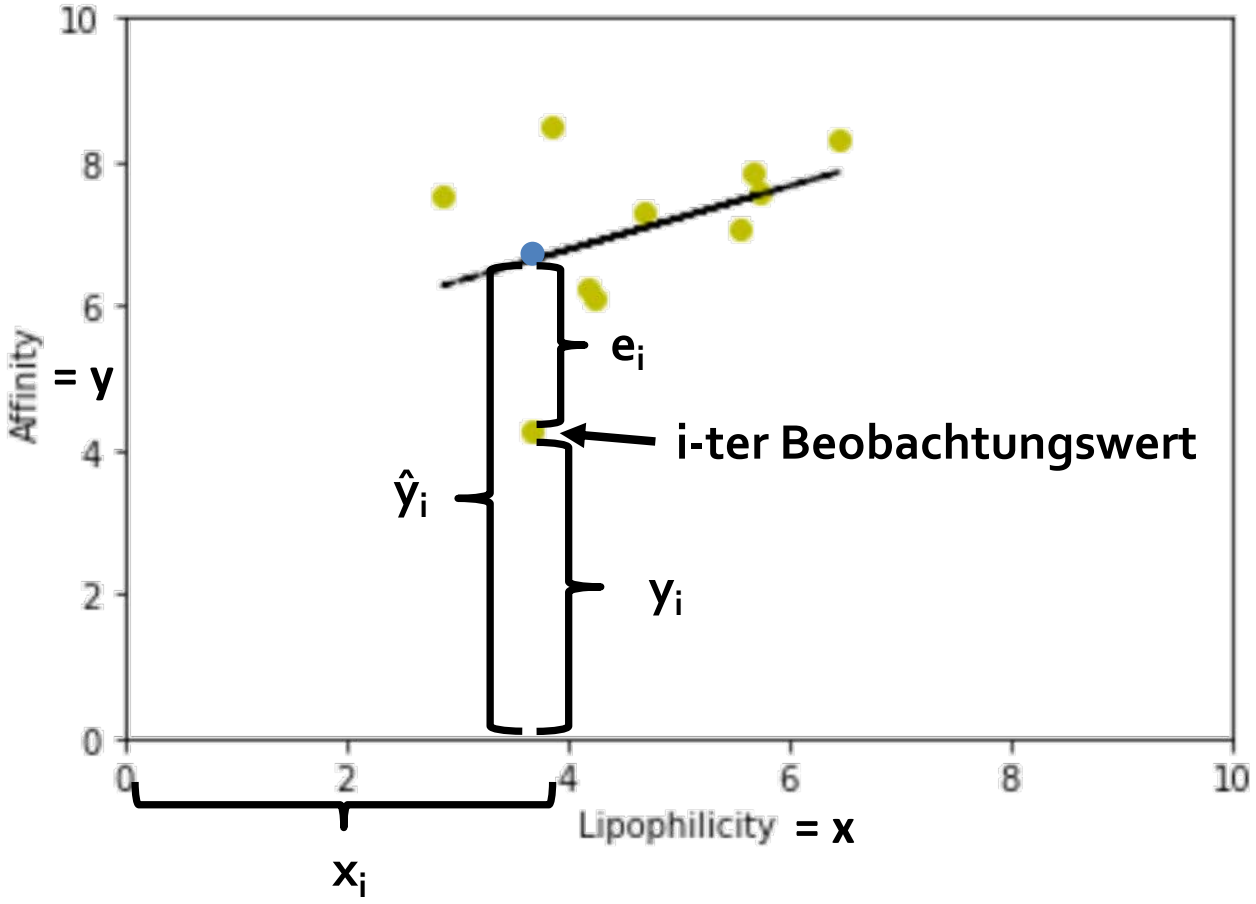


y_i Beobachtungswert

$\hat{y}_i$ Gemäß der linearen Funktion geschätzter Wert

Inspiriert von https://www.youtube.com/watch?v=osPwkFkYvNc

Geradengleichung



$$\hat{y}_i = a+b*x_i$$

a, b: zu schätzende Parameter
a: y-Achsenabschnitt
b: Steigung der linearen Funktion

Inspiriert von https://www.youtube.com/watch?v=osPwkFkYvNc

Residuen und Zielfunktion

Die Abweichungen der geschätzten von den tatsächlichen Werten der abhängigen Variable y heißen **Residuen**

$$e_i = y_i - \hat{y}_i$$

Zielfunktion der linearen Regression (Methode der kleinesten Quadrate)

$$\sum_{i=1}^{n} e_i^2 = e_1^2 + e_2^2 + \cdots + e_n^2 \longrightarrow min_{a,b}$$

n: Anzahl der Beobachtungen

Quadriert weil:

- Residuen können positiv oder negativ sein
- Größere Residuen werden stärker gewichtet

Inspiriert von https://www.youtube.com/watch?v=osPwkFkYvNc

Ermittlung der zu schätzenden Parameter

Die Schätzung der Regressionsparameter *a* und *b* erfolgt erfolgt mittels der Methode der kleinsten Quadrate, so dass die Summe der quadrierten Residuen minimal wird.

Lösung des Optimierungsproblems

y-Achsenabschnitt (*a*)

$$a = \bar{y} - b\bar{x}$$

Steigung (*b*)

Kovarianz

Korrelationskoeffizient nach Bravais-Pearson

$$b = \frac{S_{XY}}{S_X^2} = \frac{S_{XY} \cdot S_Y}{S_X^2 \cdot S_Y} = r \cdot \frac{S_Y}{S_X}$$

Varianz

Inspiriert von https://www.youtube.com/watch?v=osPwkFkYvNc

Einfache, lineare Regression: Rechenbeispiel

	Molekül1	Molekül2	Molekül3	Molekül4	Molekül5	Molekül6	Molekül7	Molekül8	Molekül9	Molekül10
Lipophilie	3,84	5,67	5,55	2,88	4,24	3,69	6,45	4,18	4,68	5,74
Affinität	8,52	7,85	7,08	7,54	6,09	4,29	8,3	6,26	7,32	7,57

$$a = \bar{y} - b\bar{x}$$

$$b = \frac{S_{XY}}{S_X^2} = \frac{S_{XY} \cdot S_Y}{S_X^2 \cdot S_Y} = r \cdot \frac{S_Y}{S_X}$$

Was benötigen wir?

1) Arithmetisches Mittel von *X*: $\bar{x}$
2) Arithmetisches Mittel von *Y*: $\bar{y}$
3) Kovarianz von *X* und *Y*: S_{XY}
4) Varianz von *X*: S^2_X

(3) $$s_{XY} = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y}) = \frac{1}{n} \cdot \left(\sum_{i=1}^{n} x_i \cdot y_i \right) - \bar{x} \cdot \bar{y}$$

(4) $$s_X^2 = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n} \cdot \left(\sum_{i=1}^{n} x_i^2 \right) - \bar{x}^2$$

Hinweis $\tilde{S}_{XY}$ und S_{XY} (sowie $\tilde{S}_X$ und S_X etc.) bedeuten dasselbe

Inspiriert von https://www.youtube.com/watch?v=osPwkFkYvNc

Lineare Regression in Python

```
x_mean = np.mean(Lipophilicity)
```

```
y_mean = np.mean(Affinity)
```

```
np.cov(Lipophilicity,Affinity)
```

```
array([[1.26237333, 0.55677333],
       [0.55677333, 1.56675111]])
```

```
S_XY = np.cov(Lipophilicity,Affinity,bias=True)[0,1]
S_XY
```

```
0.5010960000000002
```

```
S_x2 = np.var(Lipophilicity)
S_x2
```

```
1.136136
```

```
b = S_XY / S_x2
b
```

```
0.44105283170324694
```

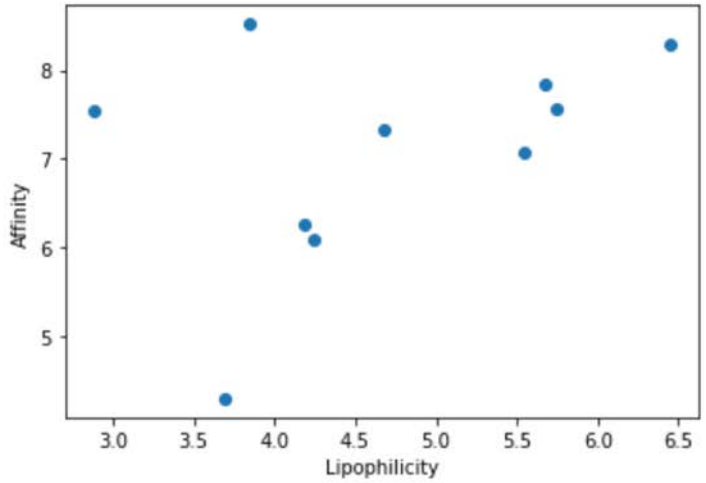
```
yachsenabschnitt = y_mean - b*x_mean
yachsenabschnitt
```

```
5.012580113648364
```

Lineare Regression in Python

```
Lipophilicity = [3.84, 5.67, 5.55, 2.88, 4.24, 3.69, 6.45, 4.18, 4.68, 5.74]
Affinity = [8.52, 7.85, 7.08, 7.54, 6.09, 4.29, 8.3, 6.26, 7.32, 7.57]
```

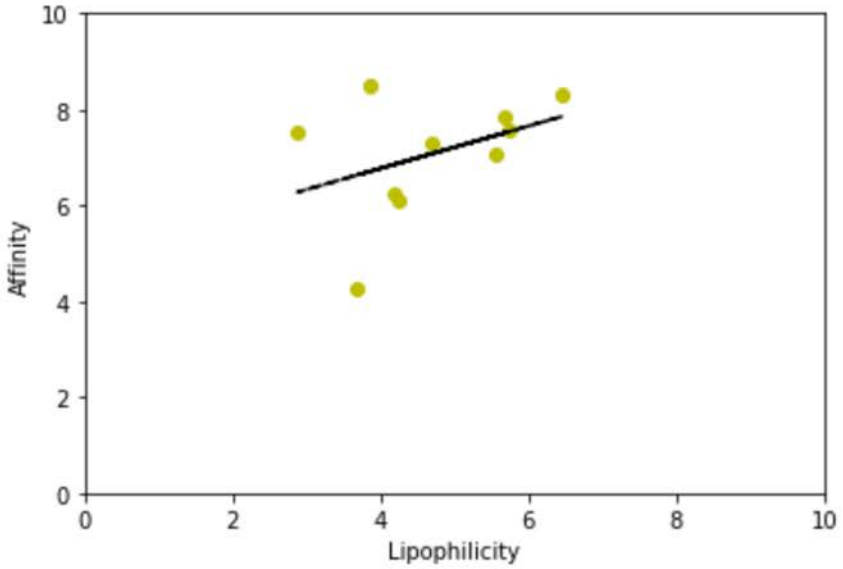
```
plt.scatter( Lipophilicity, Affinity)
plt.xlabel("Lipophilicity")
plt.ylabel("Affinity")
plt.show()
```



Lineare Regression in Python

```
coef = np.polyfit(Lipophilicity,Affinity,1)
poly1d_fn = np.poly1d(coef)
```

```
plt.plot(Lipophilicity,Affinity, 'yo', Lipophilicity, poly1d_fn(Lipophilicity), '--k')
plt.xlabel("Lipophilicity")
plt.ylabel("Affinity")
plt.xlim(0,10)
plt.ylim(0,10)
plt.show()
```



Lineare Regression in Python

```
slope, intercept, r_value, p_value, std_err = linregress(Lipophilicity,Affinity)
```

```
print("Steigung:           %1.2f" %slope)
print("y-Achsenabschnitt: %1.2f" %intercept)
print("Korrelation:        %1.2f" %r_value)
```

```
Steigung:           0.44
y-Achsenabschnitt: 5.01
Korrelation:        0.40
```

Kurze Zwischenfrage: Was bedeuten y-Achsenabschnitt und Steigung anschaulich?

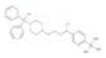
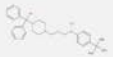
SoSe

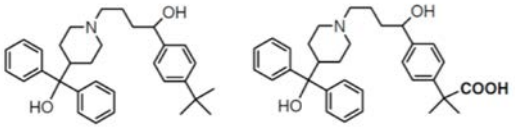
Einführung Data Science im Bereich Chemie und Chemische Biologie

Data curation

Sources of errors: Data Extraction

Unit inconsistencies is very common during data digitization

☐	 CHEMBL17157	41, Terfenadine	IC50	=	56000	nM	No Data	No Data	CHEMBL907269	Inhibition of human ERG potassium channel in HEK293 cells by patch clamp assay	cell-based format	Homo sapiens
☐	 CHEMBL17157	Terfenadine	IC50	=	56	nM	7.25	No Data	CHEMBL768686	Inhibitory activity against Potassium channel HERG	single protein format	Homo sapiens



Terfenadine **41**
hERG IC50 = 56 nM

Fexofenadine **42**
hERG IC50 = 23uM

Zhu et al. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 5507

Table 1. Comparison of the HERG Channel Affinity to That of the Intended Pharmacological Target for Several Drugs

drug	target affinity	HERG IC_{50}	comment
terfenadine	58 nM (histamine H1 K_i)	56 nM	withdrawn
astemizole	3 nM (histamine H1 K_i)	0.9 nM	withdrawn
cisapride	29 nM (serotonin 5HT4 K_i)	47 nM	withdrawn
sertindole	0.6 nM (serotonin 5HT2A K_i)	3 nM	withdrawn
thioridazine	27 nM (dopamine D2 K_i)	191 nM	black box[a]
pimozide	12 nM (dopamine D2 K_i)	18 nM	TDP[b]
grepafloxacin	up to 2.4 μM (bacterial MIC[c])	50 μM	withdrawn

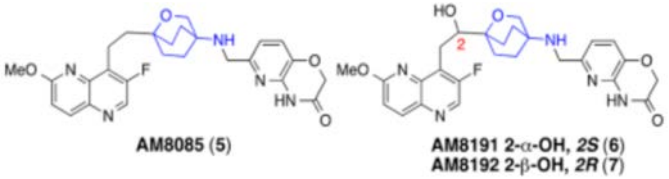
Pearlstein et al. *J. Med. Chem.* **2003**, *46*, 2017

Automation procedure: nM switched to µM

Courtesy: Pankaj Daga (Simulations Plus)

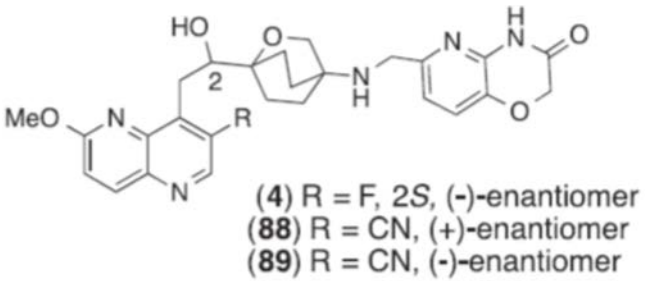
Sources of errors: Original Research Articles

Singh et al, ACS Med. Chem. Lett. **2014**, *5*, 609



agents. We evaluated hERG activity in a functional automated patch clamp assay (see Supporting Information for Methods). In this assay, AM8085 showed an IC_{50} of 0.6 μM. The 2*S*-hydroxy group of AM8191 improved the polarity and attenuated the hERG activity (IC_{50} = 18 μM). However, more than an order of magnitude attenuation of hERG activity may be required for a clinical development compound.

Singh et al, Bioorg. Med. Chem. Lett. **2015**, *25*, 2409



List	R	SaS	SaMR	Sp	Ef	Ec	Ab	Pa	hERG binding (IC_{50}, nM)	PX hERG (IC_{50}, nM)	*c*log$D_{7.4}$
4	F	0.02	.06	.05	0.5	2	0.5	8	26.00	18.00	1.9
88	CN	0.031	0.031	0.25	1	4	0.5	8	2.13	2.00	1.3
89	CN	0.063	0.031	0.25	2	4	1	16	2.58	NT	1.3

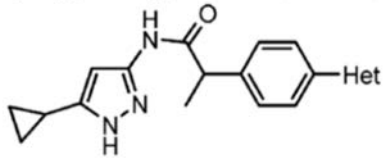
The original author made a mistake

Courtesy: Pankaj Daga (Simulations Plus)

Sources of errors: Image processing

Original publication

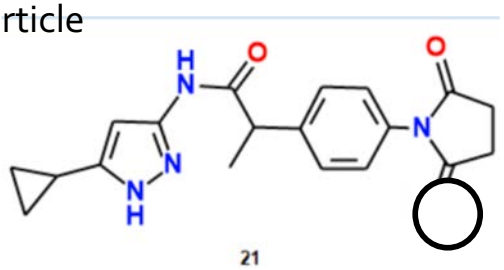
Table 3. SAR of 3-(4-Heterocycl-1-yl)phenyl-acetamido-5-cyclopropyl-1*H*-pyrazoles (**21**–**32**)



Entry	Het	α-methyl configuration	CDK2/cyclin A (IC_{50}; nM)[a]	A2780 (IC_{50}; nM)[a]	Caco-2 Permeability	Solubility (μM; buffer pH 7)	Plasma Protein Binding (%)
21		R,S	77	>10,000	Moderate	220	48
22		R,S	12	2,250	Moderate	224	74
23		R	455	13,200	Moderate	220	74
24		S	2	1,270	Moderate	>225	74
25		R,S	17	4,540	Low	201	67
16		R,S	150	6,400		222	67

Pevarello et al, *J. Med. Chem.* **2005,** *48,* 2944

The Compound **21** shows extracellular double bond in original article



21

Transcribed likewise in the database

But here an oxygen becomes lost!

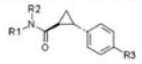
Waldman et al, *J Comput Aided Mol Des*, **2015**, *29*, 897

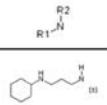
Wrong initial sketch leads to wrong structure

Courtesy: Pankaj Daga (Simulations Plus)

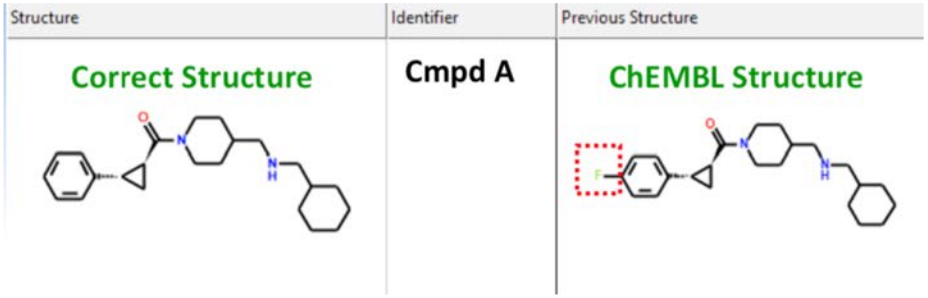
Sources of errors: Image processing

Table 1 SAR and key *in vitro* properties of phenylcyclopropylcarboxamide analogs



Compound		R3	σ1 pK_i^a	σ2 pK_i^b	Off-targets profiling	log *D* *[ACD_log D]*[c]	LLE *[cLLE]*[d]	pK_a *[ACD_pK$_a$]*[e]	Cl_{int} (rat)[f]
(±)-1		H	6.7 ± 0.38	*nt*	Ca^{2+}, 5HT2a, α1[g]	1.3	5.4	10.1	*nt*
[…]									
14		H	7.1 ± 0.22	50% inhib. at 10 μM	H3, Ca^{2+}[h]	1.1	6.0	9.3	32
15		F	7.0 ± 0.27	61% inhib. at 10 μM	Clean[h]	*[1.4]*	*[5.6]*	*[9.3]*	*nt*
16		H	7.1 ± 0.38	*nt*	Na^+, muscarinic, 5HT1a[h]	*[0.5]*	*[6.6]*	*[10.6]*	17

Valade et al, *Med Chem Commun*, **2011**, *2*, 655



The different R-Groups were combined in the wrong way

Courtesy: Pankaj Daga (Simulations Plus)