Table precise extraction from PDF file

Approach 1: Camelot

Camelot is a Python library that makes it easy to extract tables from PDF files.

Installation of dependencies

To parse tables from pdf through Camelot, dependencies **Ghostscript** and **Tkinter** must be installed. The dependencies **Ghostscript** and **Tkinter** can be installed using your system's package manager or by running their installer.

Ghostscript:

https://www.ghostscript.com/doc/current/Install.htm

Tkinter:

https://wiki.python.org/moin/TkInter

Installation of Camelot

```
$  pip install camelot-py[cv]
```

Table extraction codes

```
import camelot
import tkinter
import pandas
```

In this section, An article Energetics of Protein-Protein Interactions: Analysis of the Barnase-Barstar Interface by Single Mutations and Double Mutant Cycles downloaded from NCBI Pubmed is used as a sample file, which contains a spreadsheet of thermodynamics data in page 11.

```
tables = camelot.read_pdf('/Users/chengzihao2018/Desktop/Camelot Test/Prote
pages='11',
flavor='stream')
```

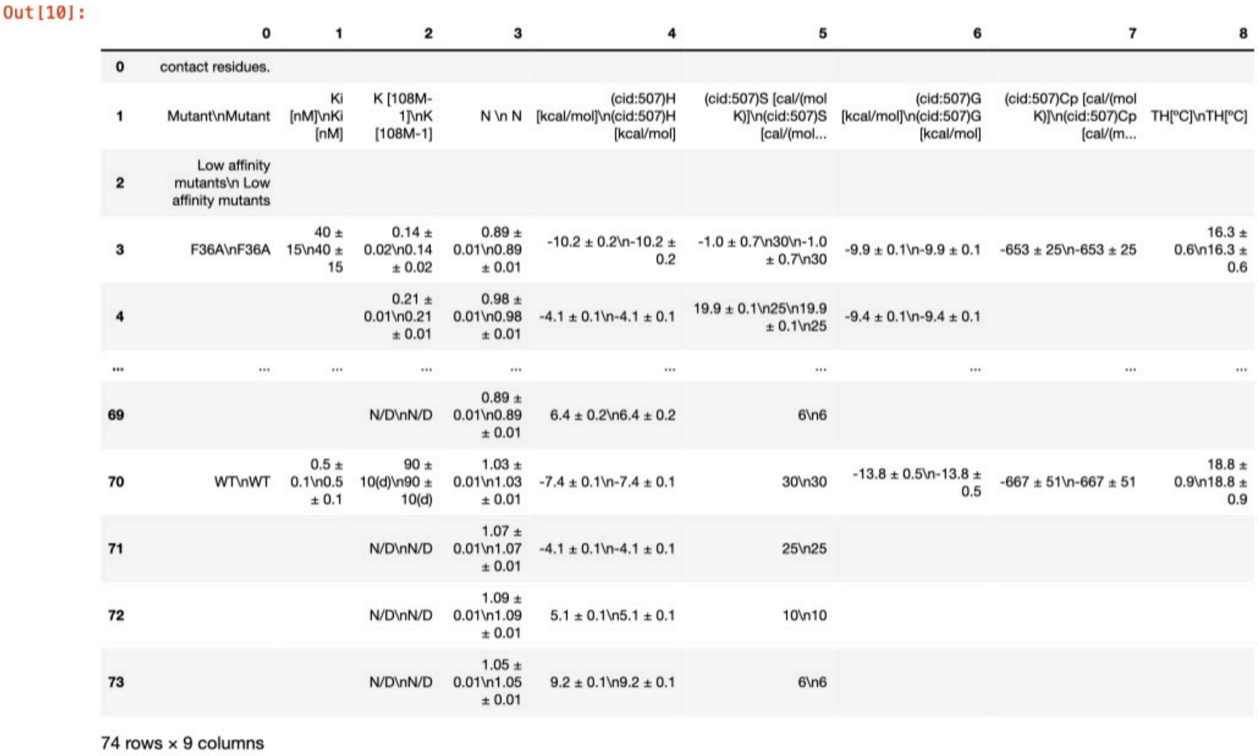
The ***Stream*** method is used to parse tables that have whitespaces between cells to simulate a table structure.The default method is ***lattice***, which is used to parse tables that have demarcated lines between cells, and can automatically parse multiple tables present on a page.

```
tables
# <TableList n=1>
```

Now, we have a TableList object called tables, which is a list of Table objects. We can get everything we need from this object.

```
tables[0]
# <Table shape=(74, 9)>
```

```
tables[0].df
```



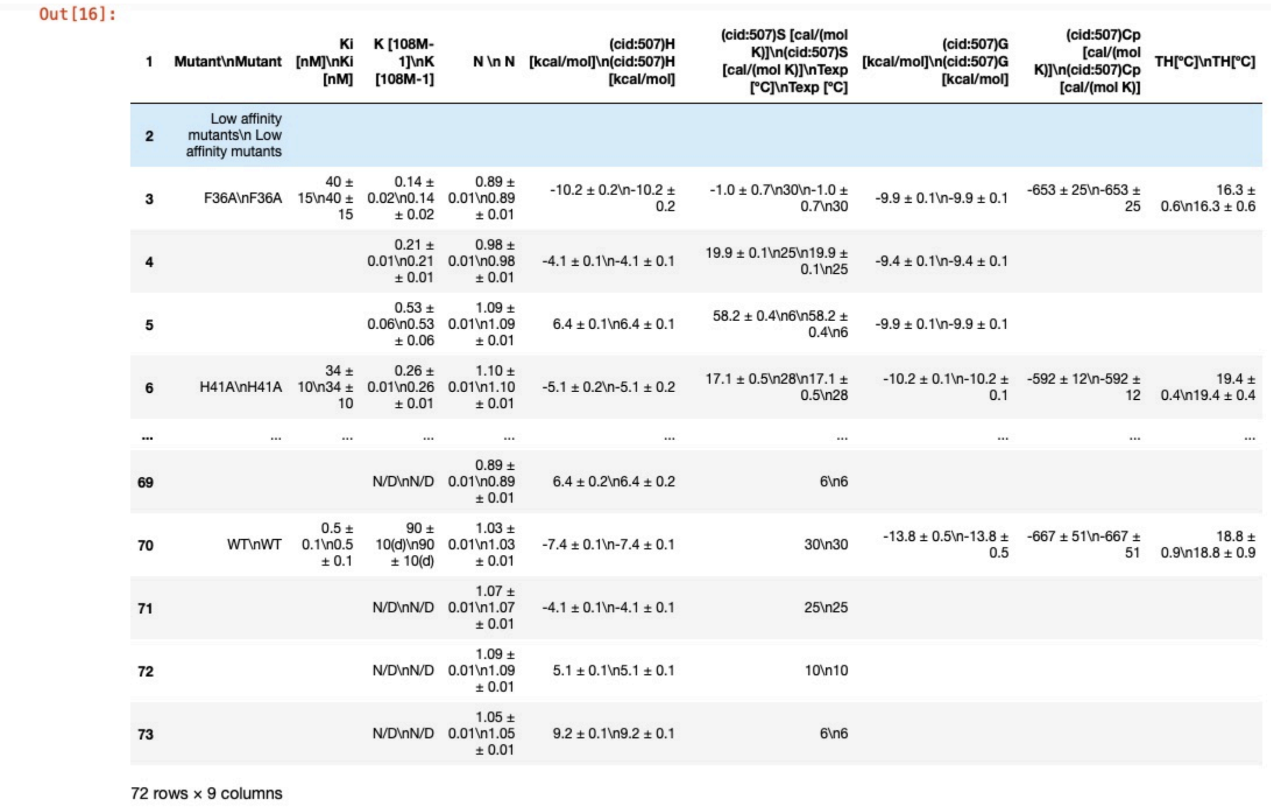
Out[10]:

	0	1	2	3	4	5	6	7	8
0	contact residues.								
1	Mutant\nMutant	Ki [nM]\nKi [nM]	K [108M-1]\nK [108M-1]	N \n N	(cid:507)H [kcal/mol]\n(cid:507)H [kcal/mol]	(cid:507)S [cal/(mol K)]\n(cid:507)S [cal/(mol...	(cid:507)G [kcal/mol]\n(cid:507)G [kcal/mol]	(cid:507)Cp [cal/(mol K)]\n(cid:507)Cp [cal/(m...	TH[°C]\nTH[°C]
2	Low affinity mutants\n Low affinity mutants								
3	F36A\nF36A	40 ± 15\n40 ± 15	0.14 ± 0.02\n0.14 ± 0.02	0.89 ± 0.01\n0.89 ± 0.01	-10.2 ± 0.2\n-10.2 ± 0.2	-1.0 ± 0.7\n30\n-1.0 ± 0.7\n30	-9.9 ± 0.1\n-9.9 ± 0.1	-653 ± 25\n-653 ± 25	16.3 ± 0.6\n16.3 ± 0.6
4			0.21 ± 0.01\n0.21 ± 0.01	0.98 ± 0.01\n0.98 ± 0.01	-4.1 ± 0.1\n-4.1 ± 0.1	19.9 ± 0.1\n25\n19.9 ± 0.1\n25	-9.4 ± 0.1\n-9.4 ± 0.1		
...	...	...	...	...	...	...	...	...	...
69			N/D\nN/D	0.89 ± 0.01\n0.89 ± 0.01	6.4 ± 0.2\n6.4 ± 0.2	6\n6			
70	WT\nWT	0.5 ± 0.1\n0.5 ± 0.1	90 ± 10(d)\n90 ± 10(d)	1.03 ± 0.01\n1.03 ± 0.01	-7.4 ± 0.1\n-7.4 ± 0.1	30\n30	-13.8 ± 0.5\n-13.8 ± 0.5	-667 ± 51\n-667 ± 51	18.8 ± 0.9\n18.8 ± 0.9
71			N/D\nN/D	1.07 ± 0.01\n1.07 ± 0.01	-4.1 ± 0.1\n-4.1 ± 0.1	25\n25			
72			N/D\nN/D	1.09 ± 0.01\n1.09 ± 0.01	5.1 ± 0.1\n5.1 ± 0.1	10\n10			
73			N/D\nN/D	1.05 ± 0.01\n1.05 ± 0.01	9.2 ± 0.1\n9.2 ± 0.1	6\n6			

74 rows × 9 columns

Now the table has been successfully extracted, but not in perfect accuracy. We can see that the cell `contact residues` on the left top is actually mistakenly parsed from the text outside the table in the article. A minor trimming to remove the unwanted row is needed.

```
df1 = df.drop([0])
new_header = df1.iloc[0] #grab the first row for the header
df2 = df1[1:] #take the data less the header row
df2.columns = new_header #set the header row as the df header
df2
```



Out[16]:

1	Mutant\nMutant	Ki [nM]\nKi [nM]	K [108M-1]\nK [108M-1]	N \n N	(cid:507)H [kcal/mol]\n(cid:507)H [kcal/mol]	(cid:507)S [cal/(mol K)]\n(cid:507)S [cal/(mol K)]\nTexp [°C]\nTexp [°C]	(cid:507)G [kcal/mol]\n(cid:507)G [kcal/mol]	(cid:507)Cp [cal/(mol K)]\n(cid:507)Cp [cal/(mol K)]	TH[°C]\nTH[°C]
2	Low affinity mutants\n Low affinity mutants								
3	F36A\nF36A	40 ± 15\n40 ± 15	0.14 ± 0.02\n0.14 ± 0.02	0.89 ± 0.01\n0.89 ± 0.01	-10.2 ± 0.2\n-10.2 ± 0.2	-1.0 ± 0.7\n30\n-1.0 ± 0.7\n30	-9.9 ± 0.1\n-9.9 ± 0.1	-653 ± 25\n-653 ± 25	16.3 ± 0.6\n16.3 ± 0.6
4			0.21 ± 0.01\n0.21 ± 0.01	0.98 ± 0.01\n0.98 ± 0.01	-4.1 ± 0.1\n-4.1 ± 0.1	19.9 ± 0.1\n25\n19.9 ± 0.1\n25	-9.4 ± 0.1\n-9.4 ± 0.1		
5			0.53 ± 0.06\n0.53 ± 0.06	1.09 ± 0.01\n1.09 ± 0.01	6.4 ± 0.1\n6.4 ± 0.1	58.2 ± 0.4\n6\n58.2 ± 0.4\n6	-9.9 ± 0.1\n-9.9 ± 0.1		
6	H41A\nH41A	34 ± 10\n34 ± 10	0.26 ± 0.01\n0.26 ± 0.01	1.10 ± 0.01\n1.10 ± 0.01	-5.1 ± 0.2\n-5.1 ± 0.2	17.1 ± 0.5\n28\n17.1 ± 0.5\n28	-10.2 ± 0.1\n-10.2 ± 0.1	-592 ± 12\n-592 ± 12	19.4 ± 0.4\n19.4 ± 0.4
...	...	...	...	...	...	...	...	...	...
69			N/D\nN/D	0.89 ± 0.01\n0.89 ± 0.01	6.4 ± 0.2\n6.4 ± 0.2	6\n6			
70	WT\nWT	0.5 ± 0.1\n0.5 ± 0.1	90 ± 10(d)\n90 ± 10(d)	1.03 ± 0.01\n1.03 ± 0.01	-7.4 ± 0.1\n-7.4 ± 0.1	30\n30	-13.8 ± 0.5\n-13.8 ± 0.5	-667 ± 51\n-667 ± 51	18.8 ± 0.9\n18.8 ± 0.9
71			N/D\nN/D	1.07 ± 0.01\n1.07 ± 0.01	-4.1 ± 0.1\n-4.1 ± 0.1	25\n25			
72			N/D\nN/D	1.09 ± 0.01\n1.09 ± 0.01	5.1 ± 0.1\n5.1 ± 0.1	10\n10			
73			N/D\nN/D	1.05 ± 0.01\n1.05 ± 0.01	9.2 ± 0.1\n9.2 ± 0.1	6\n6			

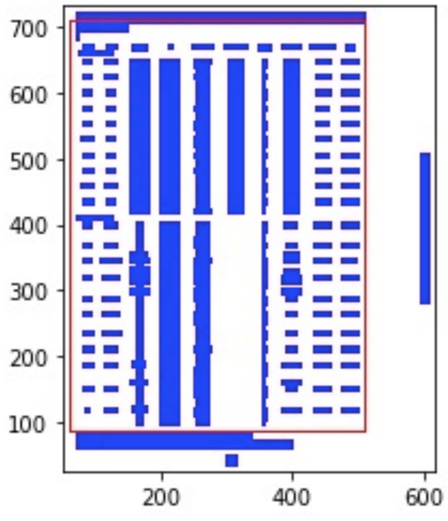
72 rows × 9 columns

Now we can see that the first row is removed.

However, for cases that have worse parsing accuracy, an alternative is the Visual Debugging function.

```
camelot.plot(tables[0], kind='contour').show()
```

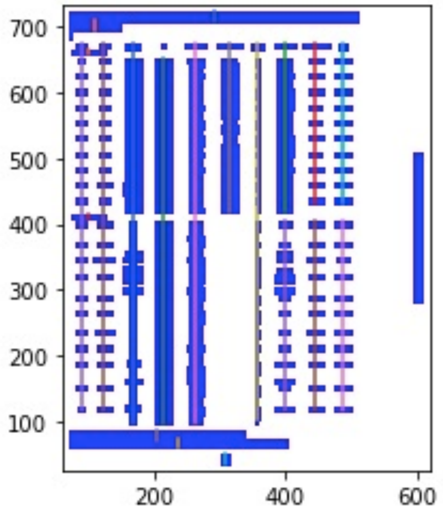
By plotting the contour of parsed tables, borders of the table can be specified:



Or:

```
camelot.plot(tables[0], kind='textedge').show()
```

Textedges can also be visualized by specifying kind='textedge'.



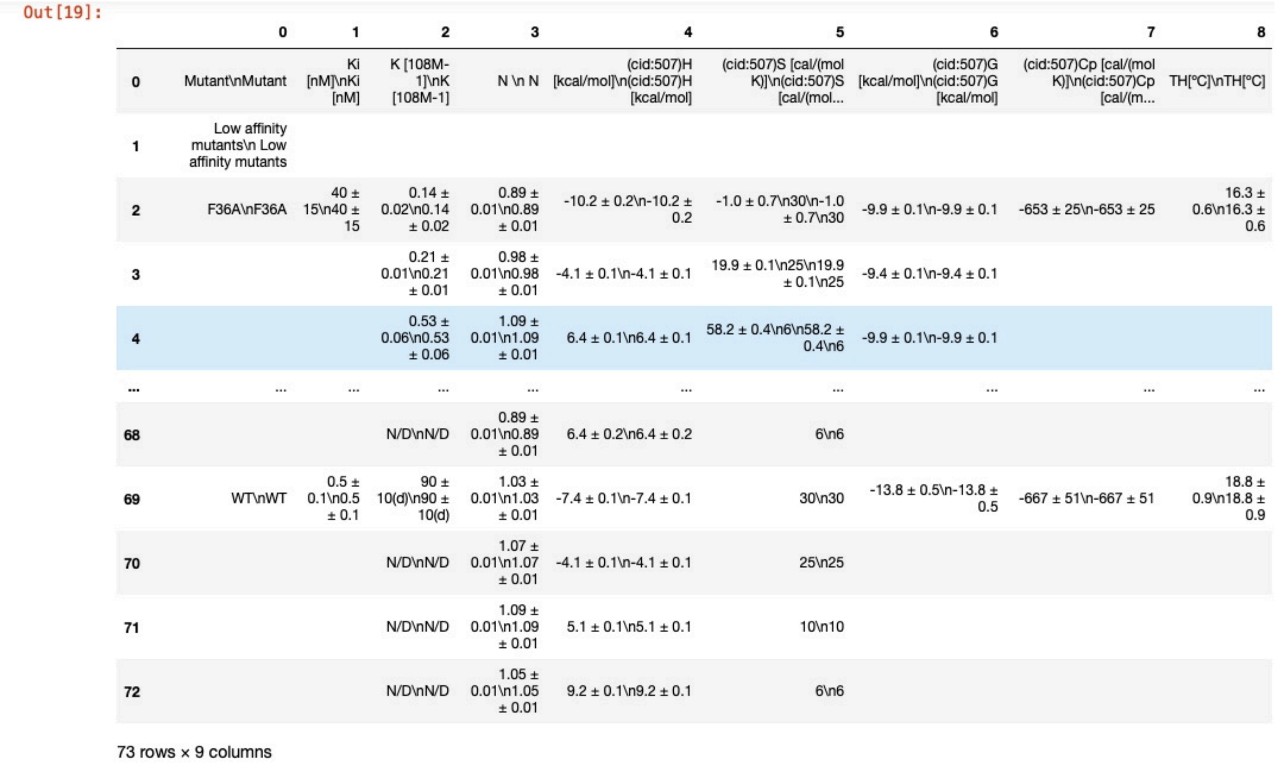
Now that you can have a better idea of the table region by visulize it in the coordinate system.

You can use the `table_areas` keyword argument to `read_pdf()` to solve for such cases. table_areas accepts strings of the form: `x1,y1,x2,y2`

(x1, y1) -> top-left and (x2, y2) -> bottom-right in PDF coordinate space.

In PDF coordinate space, the bottom-left corner of the page is the origin, with coordinates (0, 0). When `table_areas` is specified, Camelot will only analyze the specified regions to look for tables:

```
tables = camelot.read_pdf('/Users/chengzihao2018/Desktop/Camelot Test/Prote
                          pages='11',
                          flavor='stream',
                          table_areas=['5,695,500,100']
                          )
tables[0].df
```



Out[19]:

	0	1	2	3	4	5	6	7	8
0	Mutant\nMutant	Ki [nM]\nKi [nM]	K [108M-1]\nK [108M-1]	N \n N	(cid:507)H [kcal/mol]\n(cid:507)H [kcal/mol]	(cid:507)S [cal/(mol K)]\n(cid:507)S [cal/(mol...	(cid:507)G [kcal/mol]\n(cid:507)G [kcal/mol]	(cid:507)Cp [cal/(mol K)]\n(cid:507)Cp [cal/(m...	TH[°C]\nTH[°C]
1	Low affinity mutants\n Low affinity mutants								
2	F36A\nF36A	40 ± 15\n40 ± 15	0.14 ± 0.02\n0.14 ± 0.02	0.89 ± 0.01\n0.89 ± 0.01	-10.2 ± 0.2\n-10.2 ± 0.2	-1.0 ± 0.7\n30\n-1.0 ± 0.7\n30	-9.9 ± 0.1\n-9.9 ± 0.1	-653 ± 25\n-653 ± 25	16.3 ± 0.6\n16.3 ± 0.6
3			0.21 ± 0.01\n0.21 ± 0.01	0.98 ± 0.01\n0.98 ± 0.01	-4.1 ± 0.1\n-4.1 ± 0.1	19.9 ± 0.1\n25\n19.9 ± 0.1\n25	-9.4 ± 0.1\n-9.4 ± 0.1		
4			0.53 ± 0.06\n0.53 ± 0.06	1.09 ± 0.01\n1.09 ± 0.01	6.4 ± 0.1\n6.4 ± 0.1	58.2 ± 0.4\n6\n58.2 ± 0.4\n6	-9.9 ± 0.1\n-9.9 ± 0.1		
...	...	...	...	...	...	...	...	...	...
68			N/D\nN/D	0.89 ± 0.01\n0.89 ± 0.01	6.4 ± 0.2\n6.4 ± 0.2	6\n6			
69	WT\nWT	0.5 ± 0.1\n0.5 ± 0.1	90 ± 10(d)\n90 ± 10(d)	1.03 ± 0.01\n1.03 ± 0.01	-7.4 ± 0.1\n-7.4 ± 0.1	30\n30	-13.8 ± 0.5\n-13.8 ± 0.5	-667 ± 51\n-667 ± 51	18.8 ± 0.9\n18.8 ± 0.9
70			N/D\nN/D	1.07 ± 0.01\n1.07 ± 0.01	-4.1 ± 0.1\n-4.1 ± 0.1	25\n25			
71			N/D\nN/D	1.09 ± 0.01\n1.09 ± 0.01	5.1 ± 0.1\n5.1 ± 0.1	10\n10			
72			N/D\nN/D	1.05 ± 0.01\n1.05 ± 0.01	9.2 ± 0.1\n9.2 ± 0.1	6\n6			

73 rows × 9 columns

After modification, the dataframe can be exported as multiple types of file such as csv,excel,html,etc.

```
outputpath= '/Users/chengzihao2018/Desktop/Protein0**.csv'
df2.to_csv(outputpath,sep=',',index=False,header=True)
```

The screenshot of the original spreadsheet from the pdf file:

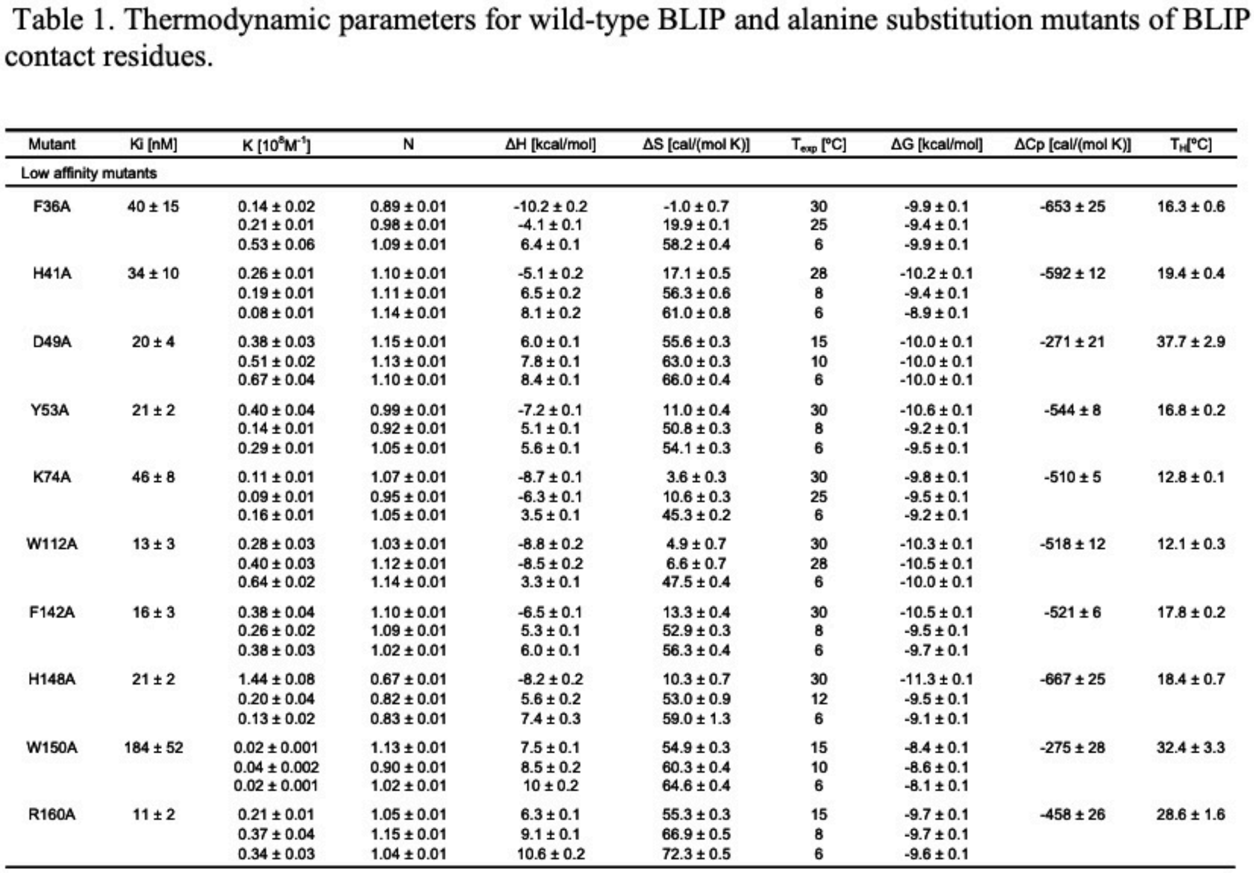


Table 1. Thermodynamic parameters for wild-type BLIP and alanine substitution mutants of BLIP contact residues.

Mutant	Ki [nM]	K [10^8M^{-1}]	N	ΔH [kcal/mol]	ΔS [cal/(mol K)]	T_{exp} [ºC]	ΔG [kcal/mol]	ΔCp [cal/(mol K)]	T_H[ºC]
Low affinity mutants									
F36A	40 ± 15	0.14 ± 0.02	0.89 ± 0.01	-10.2 ± 0.2	-1.0 ± 0.7	30	-9.9 ± 0.1	-653 ± 25	16.3 ± 0.6
		0.21 ± 0.01	0.98 ± 0.01	-4.1 ± 0.1	19.9 ± 0.1	25	-9.4 ± 0.1		
		0.53 ± 0.06	1.09 ± 0.01	6.4 ± 0.1	58.2 ± 0.4	6	-9.9 ± 0.1		
H41A	34 ± 10	0.26 ± 0.01	1.10 ± 0.01	-5.1 ± 0.2	17.1 ± 0.5	28	-10.2 ± 0.1	-592 ± 12	19.4 ± 0.4
		0.19 ± 0.01	1.11 ± 0.01	6.5 ± 0.2	56.3 ± 0.6	8	-9.4 ± 0.1		
		0.08 ± 0.01	1.14 ± 0.01	8.1 ± 0.2	61.0 ± 0.8	6	-8.9 ± 0.1		
D49A	20 ± 4	0.38 ± 0.03	1.15 ± 0.01	6.0 ± 0.1	55.6 ± 0.3	15	-10.0 ± 0.1	-271 ± 21	37.7 ± 2.9
		0.51 ± 0.02	1.13 ± 0.01	7.8 ± 0.1	63.0 ± 0.3	10	-10.0 ± 0.1		
		0.67 ± 0.04	1.10 ± 0.01	8.4 ± 0.1	66.0 ± 0.4	6	-10.0 ± 0.1		
Y53A	21 ± 2	0.40 ± 0.04	0.99 ± 0.01	-7.2 ± 0.1	11.0 ± 0.4	30	-10.6 ± 0.1	-544 ± 8	16.8 ± 0.2
		0.14 ± 0.01	0.92 ± 0.01	5.1 ± 0.1	50.8 ± 0.3	8	-9.2 ± 0.1		
		0.29 ± 0.01	1.05 ± 0.01	5.6 ± 0.1	54.1 ± 0.3	6	-9.5 ± 0.1		
K74A	46 ± 8	0.11 ± 0.01	1.07 ± 0.01	-8.7 ± 0.1	3.6 ± 0.3	30	-9.8 ± 0.1	-510 ± 5	12.8 ± 0.1
		0.09 ± 0.01	0.95 ± 0.01	-6.3 ± 0.1	10.6 ± 0.3	25	-9.5 ± 0.1		
		0.16 ± 0.01	1.05 ± 0.01	3.5 ± 0.1	45.3 ± 0.2	6	-9.2 ± 0.1		
W112A	13 ± 3	0.28 ± 0.03	1.03 ± 0.01	-8.8 ± 0.2	4.9 ± 0.7	30	-10.3 ± 0.1	-518 ± 12	12.1 ± 0.3
		0.40 ± 0.03	1.12 ± 0.01	-8.5 ± 0.2	6.6 ± 0.7	28	-10.5 ± 0.1		
		0.64 ± 0.02	1.14 ± 0.01	3.3 ± 0.1	47.5 ± 0.4	6	-10.0 ± 0.1		
F142A	16 ± 3	0.38 ± 0.04	1.10 ± 0.01	-6.5 ± 0.1	13.3 ± 0.4	30	-10.5 ± 0.1	-521 ± 6	17.8 ± 0.2
		0.26 ± 0.02	1.09 ± 0.01	5.3 ± 0.1	52.9 ± 0.3	8	-9.5 ± 0.1		
		0.38 ± 0.03	1.02 ± 0.01	6.0 ± 0.1	56.3 ± 0.4	6	-9.7 ± 0.1		
H148A	21 ± 2	1.44 ± 0.08	0.67 ± 0.01	-8.2 ± 0.2	10.3 ± 0.7	30	-11.3 ± 0.1	-667 ± 25	18.4 ± 0.7
		0.20 ± 0.04	0.82 ± 0.01	5.6 ± 0.2	53.0 ± 0.9	12	-9.5 ± 0.1		
		0.13 ± 0.02	0.83 ± 0.01	7.4 ± 0.3	59.0 ± 1.3	6	-9.1 ± 0.1		
W150A	184 ± 52	0.02 ± 0.001	1.13 ± 0.01	7.5 ± 0.1	54.9 ± 0.3	15	-8.4 ± 0.1	-275 ± 28	32.4 ± 3.3
		0.04 ± 0.002	0.90 ± 0.01	8.5 ± 0.2	60.3 ± 0.4	10	-8.6 ± 0.1		
		0.02 ± 0.001	1.02 ± 0.01	10 ± 0.2	64.6 ± 0.4	6	-8.1 ± 0.1		
R160A	11 ± 2	0.21 ± 0.01	1.05 ± 0.01	6.3 ± 0.1	55.3 ± 0.3	15	-9.7 ± 0.1	-458 ± 26	28.6 ± 1.6
		0.37 ± 0.04	1.15 ± 0.01	9.1 ± 0.1	66.9 ± 0.5	8	-9.7 ± 0.1		
		0.34 ± 0.03	1.04 ± 0.01	10.6 ± 0.2	72.3 ± 0.5	6	-9.6 ± 0.1		

Approach 2: Excalibur

Excalibur is a web interface to extract tabular data from PDFs, written in Python 3. It is powered by ***Camelot***.

Installation and Using Excalibur

You need to install ***Ghostscript*** before moving forward.

```
$ pip install excalibur-py
```

Then you need to initialize the metadata database using:

```
$ excalibur initdb
```

And then start the webserver using:

```
$ excalibur webserver
```

Now you can go to http://localhost:5000 and start extracting tabular data from your PDFs:

1.Upload a PDF and enter the page numbers you want to extract tables from.

2.Go to each page and select the table by drawing a box around it. (You can choose to skip this step since Excalibur can automatically detect tables on its own.

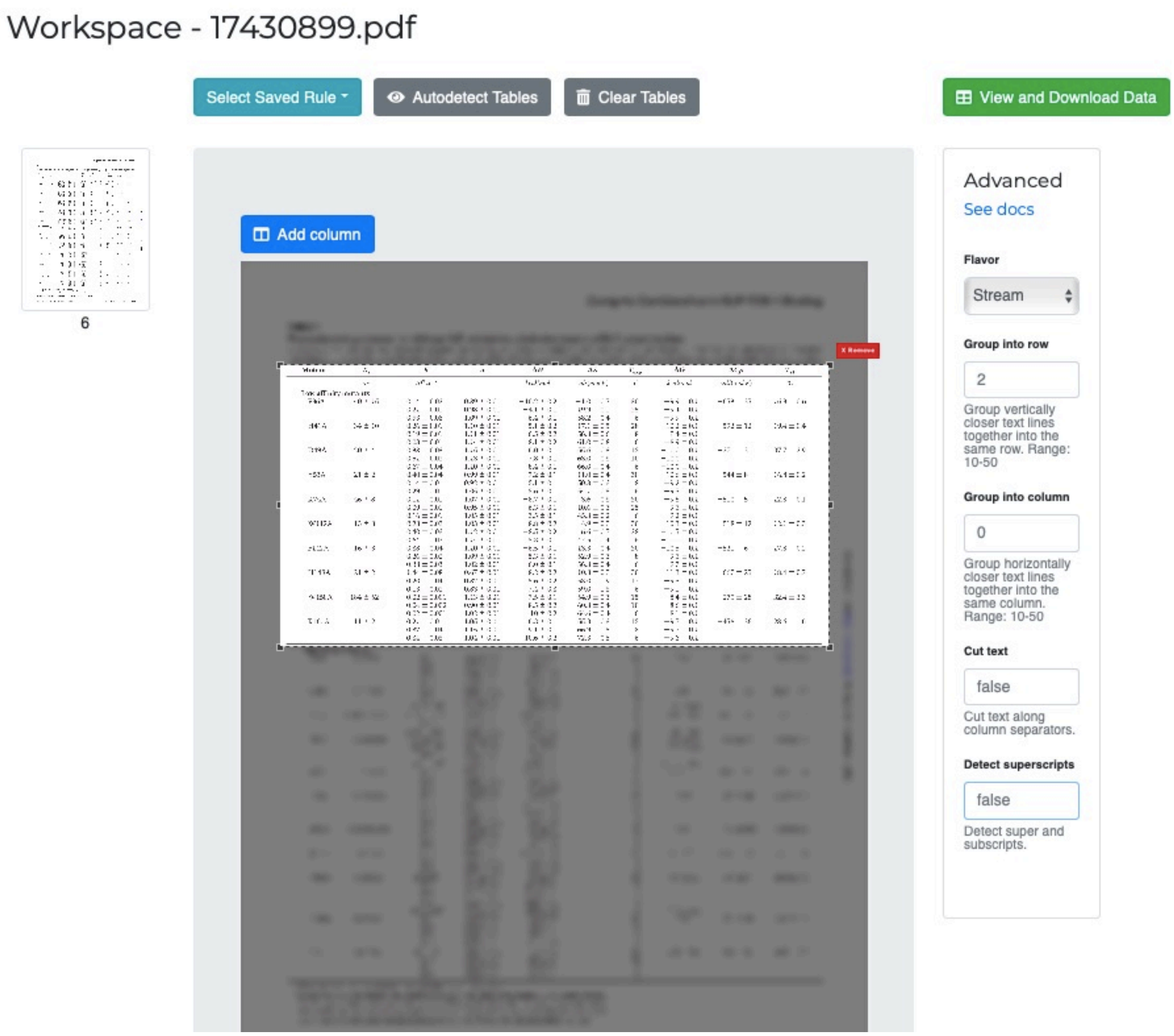
3.Click on "Autodetect tables" to see what Excalibur sees.)

4.Choose a flavor (Lattice or Stream) from "Advanced".

5.Click on "View and download data" to see the extracted tables.

6.Select your favorite format (CSV/Excel/JSON/HTML) and click on "Download"!

Example



Workspace - 17430899.pdf

Advanced

Flavor: Stream

Group into row: 2
Group vertically closer text lines together into the same row. Range: 10-50

Group into column: 0
Group horizontally closer text lines together into the same column. Range: 10-50

Cut text: false
Cut text along column separators.

Detect superscripts: false
Detect super and subscripts.

Note: by setting the ***Detect superscripts*** to be `True`, superscripts will be ***flagged*** by CID fonts, such as `<s>` for subscript, but not correctly displayed.

Here is the comparison between the original table and the extracted table:

Original table screenshot from the pdf:

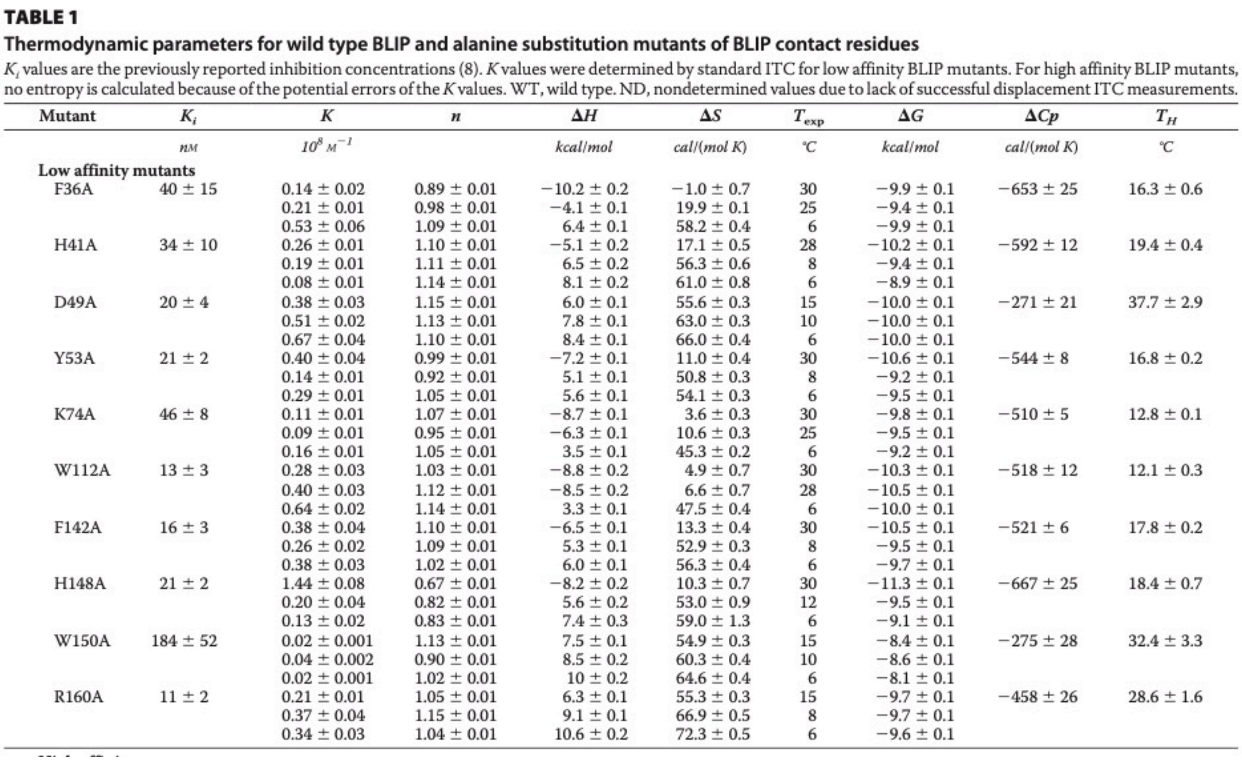


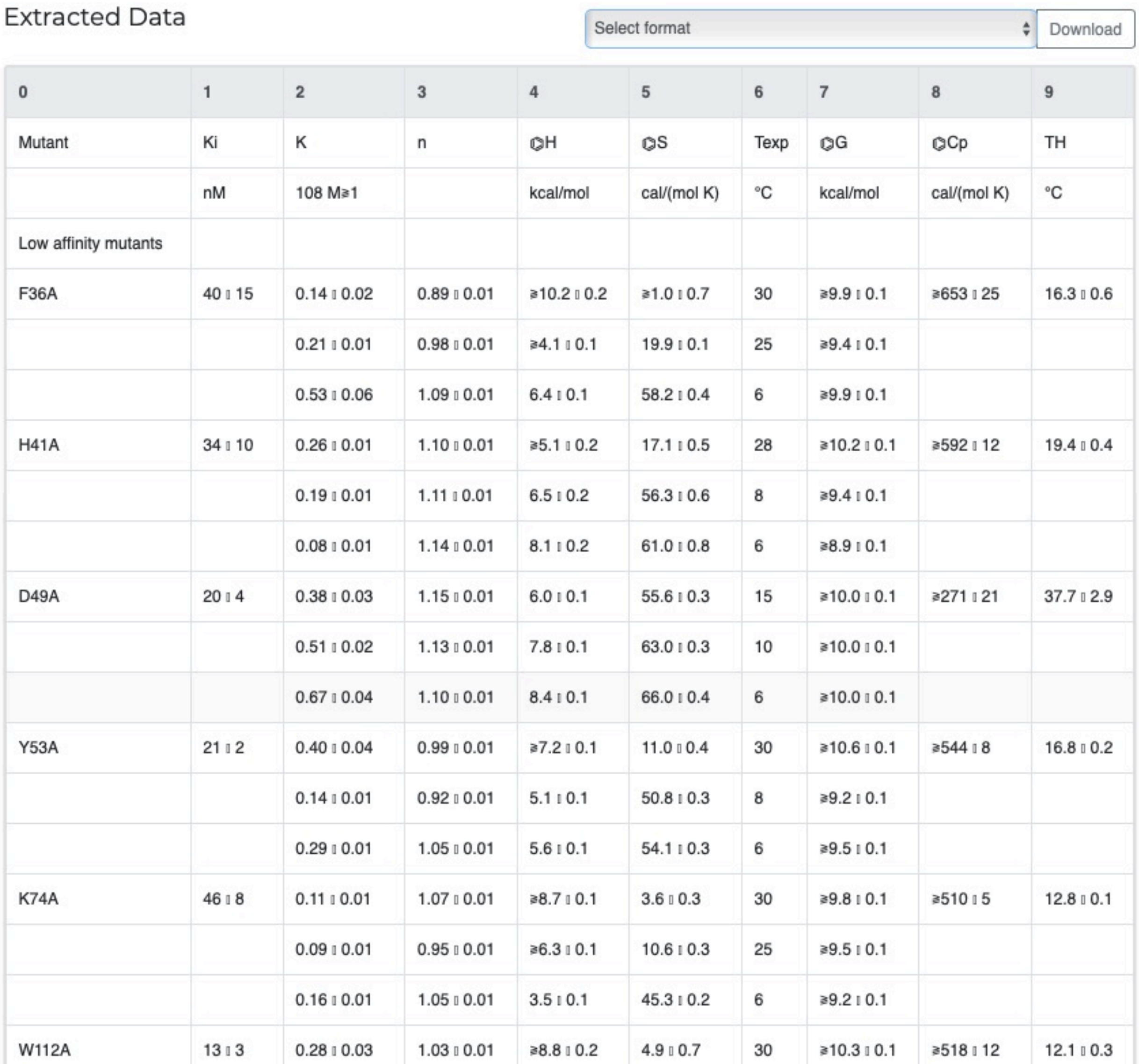
TABLE 1

Thermodynamic parameters for wild type BLIP and alanine substitution mutants of BLIP contact residues

K_i values are the previously reported inhibition concentrations (8). K values were determined by standard ITC for low affinity BLIP mutants. For high affinity BLIP mutants, no entropy is calculated because of the potential errors of the K values. WT, wild type. ND, nondetermined values due to lack of successful displacement ITC measurements.

Mutant	K_i	K	n	ΔH	ΔS	T_{exp}	ΔG	ΔCp	T_H
	nM	10^8 M^{-1}		*kcal/mol*	*cal/(mol K)*	°C	*kcal/mol*	*cal/(mol K)*	°C
Low affinity mutants									
F36A	40 ± 15	0.14 ± 0.02	0.89 ± 0.01	−10.2 ± 0.2	−1.0 ± 0.7	30	−9.9 ± 0.1	−653 ± 25	16.3 ± 0.6
		0.21 ± 0.01	0.98 ± 0.01	−4.1 ± 0.1	19.9 ± 0.1	25	−9.4 ± 0.1		
		0.53 ± 0.06	1.09 ± 0.01	6.4 ± 0.1	58.2 ± 0.4	6	−9.9 ± 0.1		
H41A	34 ± 10	0.26 ± 0.01	1.10 ± 0.01	−5.1 ± 0.2	17.1 ± 0.5	28	−10.2 ± 0.1	−592 ± 12	19.4 ± 0.4
		0.19 ± 0.01	1.11 ± 0.01	6.5 ± 0.2	56.3 ± 0.6	8	−9.4 ± 0.1		
		0.08 ± 0.01	1.14 ± 0.01	8.1 ± 0.2	61.0 ± 0.8	6	−8.9 ± 0.1		
D49A	20 ± 4	0.38 ± 0.03	1.15 ± 0.01	6.0 ± 0.1	55.6 ± 0.3	15	−10.0 ± 0.1	−271 ± 21	37.7 ± 2.9
		0.51 ± 0.02	1.13 ± 0.01	7.8 ± 0.1	63.0 ± 0.3	10	−10.0 ± 0.1		
		0.67 ± 0.04	1.10 ± 0.01	8.4 ± 0.1	66.0 ± 0.4	6	−10.0 ± 0.1		
Y53A	21 ± 2	0.40 ± 0.04	0.99 ± 0.01	−7.2 ± 0.1	11.0 ± 0.4	30	−10.6 ± 0.1	−544 ± 8	16.8 ± 0.2
		0.14 ± 0.01	0.92 ± 0.01	5.1 ± 0.1	50.8 ± 0.3	8	−9.2 ± 0.1		
		0.29 ± 0.01	1.05 ± 0.01	5.6 ± 0.1	54.1 ± 0.3	6	−9.5 ± 0.1		
K74A	46 ± 8	0.11 ± 0.01	1.07 ± 0.01	−8.7 ± 0.1	3.6 ± 0.3	30	−9.8 ± 0.1	−510 ± 5	12.8 ± 0.1
		0.09 ± 0.01	0.95 ± 0.01	−6.3 ± 0.1	10.6 ± 0.3	25	−9.5 ± 0.1		
		0.16 ± 0.01	1.05 ± 0.01	3.5 ± 0.1	45.3 ± 0.2	6	−9.2 ± 0.1		
W112A	13 ± 3	0.28 ± 0.03	1.03 ± 0.01	−8.8 ± 0.2	4.9 ± 0.7	30	−10.3 ± 0.1	−518 ± 12	12.1 ± 0.3
		0.40 ± 0.03	1.12 ± 0.01	−8.5 ± 0.2	6.6 ± 0.7	28	−10.5 ± 0.1		
		0.64 ± 0.02	1.14 ± 0.01	3.3 ± 0.1	47.5 ± 0.4	6	−10.0 ± 0.1		
F142A	16 ± 3	0.38 ± 0.04	1.10 ± 0.01	−6.5 ± 0.1	13.3 ± 0.4	30	−10.5 ± 0.1	−521 ± 6	17.8 ± 0.2
		0.26 ± 0.02	1.09 ± 0.01	5.3 ± 0.1	52.9 ± 0.3	8	−9.5 ± 0.1		
		0.38 ± 0.03	1.02 ± 0.01	6.0 ± 0.1	56.3 ± 0.4	6	−9.7 ± 0.1		
H148A	21 ± 2	1.44 ± 0.08	0.67 ± 0.01	−8.2 ± 0.2	10.3 ± 0.7	30	−11.3 ± 0.1	−667 ± 25	18.4 ± 0.7
		0.20 ± 0.04	0.82 ± 0.01	5.6 ± 0.2	53.0 ± 0.9	12	−9.5 ± 0.1		
		0.13 ± 0.02	0.83 ± 0.01	7.4 ± 0.3	59.0 ± 1.3	6	−9.1 ± 0.1		
W150A	184 ± 52	0.02 ± 0.001	1.13 ± 0.01	7.5 ± 0.1	54.9 ± 0.3	15	−8.4 ± 0.1	−275 ± 28	32.4 ± 3.3
		0.04 ± 0.002	0.90 ± 0.01	8.5 ± 0.2	60.3 ± 0.4	10	−8.6 ± 0.1		
		0.02 ± 0.001	1.02 ± 0.01	10 ± 0.2	64.6 ± 0.4	6	−8.1 ± 0.1		
R160A	11 ± 2	0.21 ± 0.01	1.05 ± 0.01	6.3 ± 0.1	55.3 ± 0.3	15	−9.7 ± 0.1	−458 ± 26	28.6 ± 1.6
		0.37 ± 0.04	1.15 ± 0.01	9.1 ± 0.1	66.9 ± 0.5	8	−9.7 ± 0.1		
		0.34 ± 0.03	1.04 ± 0.01	10.6 ± 0.2	72.3 ± 0.5	6	−9.6 ± 0.1		

Extracted .csv file:



Extracted Data

0	1	2	3	4	5	6	7	8	9
Mutant	Ki	K	n	⌬H	⌬S	Texp	⌬G	⌬Cp	TH
	nM	108 M≥1		kcal/mol	cal/(mol K)	°C	kcal/mol	cal/(mol K)	°C
Low affinity mutants									
F36A	40 ▯ 15	0.14 ▯ 0.02	0.89 ▯ 0.01	≥10.2 ▯ 0.2	≥1.0 ▯ 0.7	30	≥9.9 ▯ 0.1	≥653 ▯ 25	16.3 ▯ 0.6
		0.21 ▯ 0.01	0.98 ▯ 0.01	≥4.1 ▯ 0.1	19.9 ▯ 0.1	25	≥9.4 ▯ 0.1		
		0.53 ▯ 0.06	1.09 ▯ 0.01	6.4 ▯ 0.1	58.2 ▯ 0.4	6	≥9.9 ▯ 0.1		
H41A	34 ▯ 10	0.26 ▯ 0.01	1.10 ▯ 0.01	≥5.1 ▯ 0.2	17.1 ▯ 0.5	28	≥10.2 ▯ 0.1	≥592 ▯ 12	19.4 ▯ 0.4
		0.19 ▯ 0.01	1.11 ▯ 0.01	6.5 ▯ 0.2	56.3 ▯ 0.6	8	≥9.4 ▯ 0.1		
		0.08 ▯ 0.01	1.14 ▯ 0.01	8.1 ▯ 0.2	61.0 ▯ 0.8	6	≥8.9 ▯ 0.1		
D49A	20 ▯ 4	0.38 ▯ 0.03	1.15 ▯ 0.01	6.0 ▯ 0.1	55.6 ▯ 0.3	15	≥10.0 ▯ 0.1	≥271 ▯ 21	37.7 ▯ 2.9
		0.51 ▯ 0.02	1.13 ▯ 0.01	7.8 ▯ 0.1	63.0 ▯ 0.3	10	≥10.0 ▯ 0.1		
		0.67 ▯ 0.04	1.10 ▯ 0.01	8.4 ▯ 0.1	66.0 ▯ 0.4	6	≥10.0 ▯ 0.1		
Y53A	21 ▯ 2	0.40 ▯ 0.04	0.99 ▯ 0.01	≥7.2 ▯ 0.1	11.0 ▯ 0.4	30	≥10.6 ▯ 0.1	≥544 ▯ 8	16.8 ▯ 0.2
		0.14 ▯ 0.01	0.92 ▯ 0.01	5.1 ▯ 0.1	50.8 ▯ 0.3	8	≥9.2 ▯ 0.1		
		0.29 ▯ 0.01	1.05 ▯ 0.01	5.6 ▯ 0.1	54.1 ▯ 0.3	6	≥9.5 ▯ 0.1		
K74A	46 ▯ 8	0.11 ▯ 0.01	1.07 ▯ 0.01	≥8.7 ▯ 0.1	3.6 ▯ 0.3	30	≥9.8 ▯ 0.1	≥510 ▯ 5	12.8 ▯ 0.1
		0.09 ▯ 0.01	0.95 ▯ 0.01	≥6.3 ▯ 0.1	10.6 ▯ 0.3	25	≥9.5 ▯ 0.1		
		0.16 ▯ 0.01	1.05 ▯ 0.01	3.5 ▯ 0.1	45.3 ▯ 0.2	6	≥9.2 ▯ 0.1		
W112A	13 ▯ 3	0.28 ▯ 0.03	1.03 ▯ 0.01	≥8.8 ▯ 0.2	4.9 ▯ 0.7	30	≥10.3 ▯ 0.1	≥518 ▯ 12	12.1 ▯ 0.3

According to the comparison, we can clearly see that some special symbols such as delta symbol and Plus-minus sign cannot be correctly parsed.

The issue occurs in most of the cases that contain such symbols.

I also notice that in some pdf files, the delta symbol can be displayed correctly, however, it can also sometimes be transfered into character `D` or ***benzene ring*** symbol. In some other output csv files, Plus-minus can be displayed correctly, or can be mistakenly transfered to `6`.

The possible explanation could be that PDF has its encoding system of `binary` ,while text file or csv file use `utf-8` instead.