

# **CLiDE: improved accuracy in extracting chemical structure data from documents**

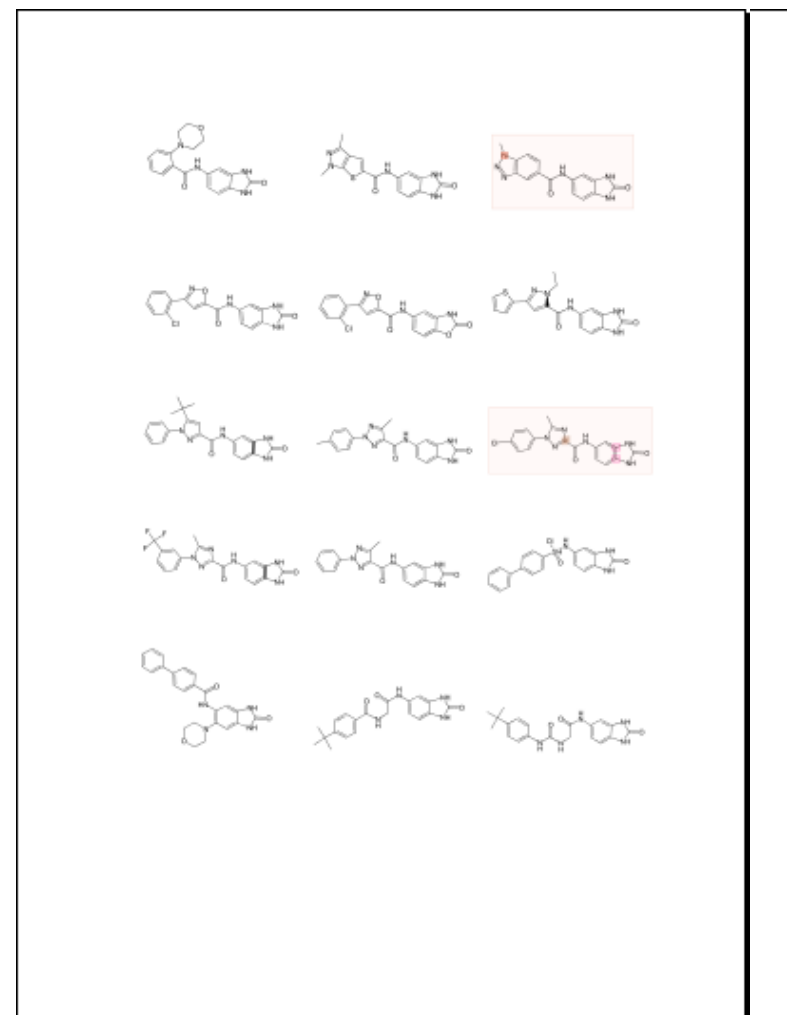
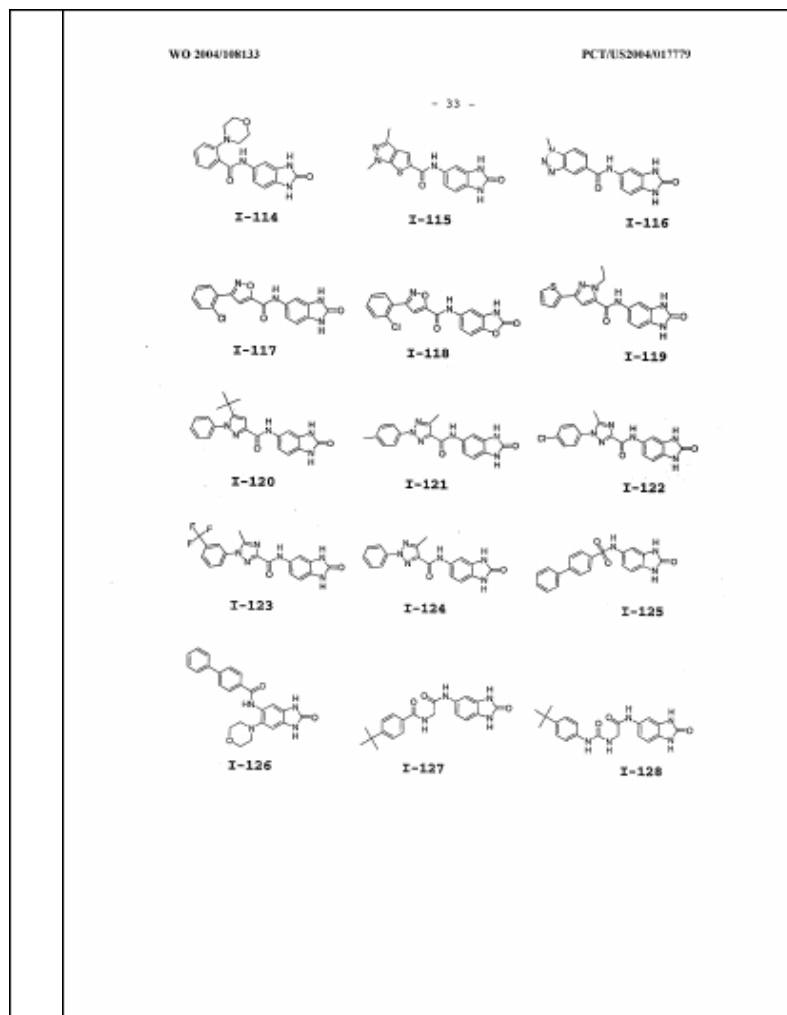
**Aniko T. Valko**  
**Keymodule Ltd.**

Peter Johnson

Vilmos A. Valko

# What is CLiDE for?

- 1) Identification of chemical structure diagrams depicted in document pages
- 2) Interpreting chemical structure diagrams
- 3) Indication of interpretation errors



# Flavours of CLiDE

CLiDE is released in three variants, designed for individual user needs



## CLiDE Standard

LogBB-2005.pdf - CLiDE Standard

File Edit View Go Tools Help

Previous Next 1 of 8 125% AL Text Select Lasso Extract Cancel Extraction

Structures and corresponding logBB values of the compounds in the Training Set:

No	Name	LogBB	No	Name	LogBB	No	Name	LogBB
1	Cimetidine	-1.42	2	ICI17148	-0.04	3	Icotidine	-2.00
4	SK&F93319	-1.30	5	Lapitidine	-1.06	6	Clonidine	0.11
7	Mepyramine	0.49	8	Imipramine	0.83	9	Ranitidine	-1.23
10	Tiotidine	-2.15						
13	BBCPD	-0.12						
16	BBCPD	-1.57						
19	BBCPD	-0.73						
22	BBCPD18	-0.27	23	BBCPD19	-0.28	24	BBCPD20	-0.46
25	BBCPD21	-0.24	26	BBCPD22	-0.02	27	BBCPD23	0.69
28	BBCPD24	0.44	29	Zolantidine	0.14	30	BBCPD26	0.22

Extracted Structure dialog box:

Edit Save Options Close



## CLiDE Professional

LogBB-2005.pdf - CLiDE Professional

File Edit View Go Tools Help

Previous Next 1 of 9 125% AL Text Select Lasso Extract Extract Range Cancel Extraction

Structures and corresponding logBB values of the

No	Name	LogBB	No	Name	LogBB
1	Cimetidine	-1.42	2	ICI17148	
4	SK&F93319	-1.30	5	Lapitidine	
7	Mepyramine	0.49	8	Imipramine	
10	Tiotidine	-0.82	11	BBCPD10	
13	BBCPD12	-0.67	14	BBCPD13	
16	BBCPD15	-0.18	17	BBCPD57	
19	BBCPD58	-1.54	20	BBCPD17	
22	BBCPD18	-0.27	23	BBCPD19	
25	BBCPD21	-0.24	26	BBCPD22	
28	BBCPD24	0.44	29	Zolantidine	

Remove Split

Completed extracting structures from page 1



## CLiDE Batch

CLiDE Batch

```
..\Program Files (x86)\Keymodule\Clide Batch\clide-batch.bat examples\CrossingBonds.bmp
```

CLiDE Batch - Chemical Literature Data Extraction Batch  
Converts chemical diagrams to connection tables

LICENSE STATUS:  
License granted

OPTIONS:  
Write log file: no  
Export formats: sdf  
Expand superatoms: no  
Enumerate generics: no  
Interpret generics: yes  
Indicate recognition faults: yes  
Indicate valence violations: yes  
Export in 'per-page' mode: no  
Create empty output: no  
Input file: "C:\Program Files (x86)\Keymodule\Clide Batch\examples\CrossingBonds.bmp"

Loading document from 'C:\Program Files (x86)\Keymodule\Clide Batch\examples\CrossingBonds.bmp'...  
Document loaded successfully

Processing...

Saving results in 'sdf' format...  
Saved 'C:\Program Files (x86)\Keymodule\Clide Batch\examples\CrossingBonds.sdf'  
Finished saving results in 'sdf' format

Total execution time: 0 sec  
C:\Program Files (x86)\Keymodule\Clide Batch


# Developments in CLiDE

1991

1998

2006

present

 <b>UNIVERSITY OF LEEDS</b>  UK	<b>SimBioSys Inc.</b>  Canada	<b>Keymodule Ltd.</b>  UK
---	-------------------------------------	---------------------------------

Research project under the supervision of Prof. Peter Johnson

Commercialization:

- CLiDE Full
- CLiDE Lite

Source code review

Commercialization

Improvements to the structure extraction:

- auto-correction of OCR errors
- chemical formula parsing
- redesign of connection table construction
- redesign of extraction of dashed and wavy bonds
- improved noise detection and filtering

# Improvement of accuracy

1) Publicly available benchmark sets of images each of which depicts one structure diagrams

<b>USPTO</b>	<b>59.84%</b>	→	<b>93.81%</b>
<b>Maybridge</b>	<b>75.66%</b>	→	<b>90.78%</b>
<b>TREC-CHEM 2011 training</b>	<b>57.96%</b>	→	<b>92.84%</b>
<b>TREC-CHEM 2011 topics</b>	<b>57.84%</b>	→	<b>90.90%</b>

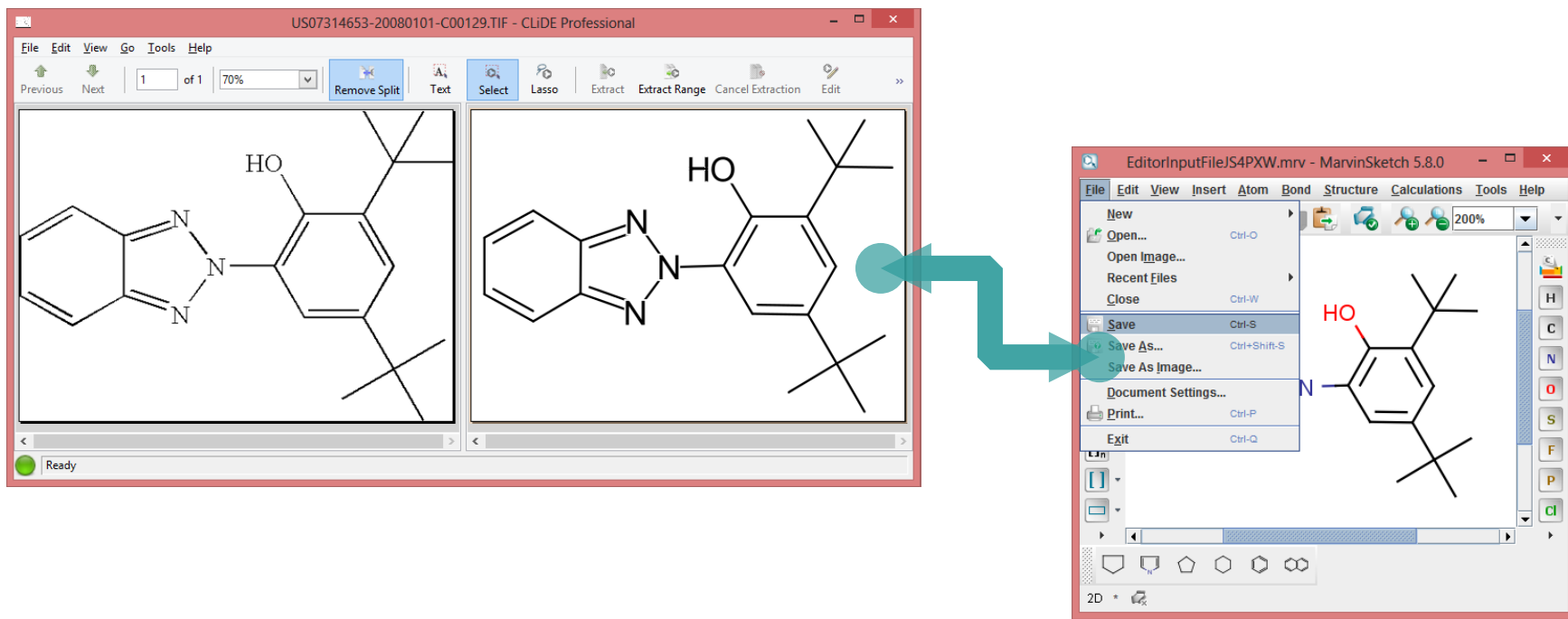
2) Non-Markush structures of two patents

<b>WO2008099019</b>	<b>12.13%</b>	→	<b>98.20%</b>
<b>US6410540</b>	<b>22.02%</b>	→	<b>97.71%</b>

Accuracy rate: the percentage of images that were correctly processed by CLiDE

# Collaboration with ChemAxon

## 1) MarvinSketch integrated with CLiDE Standard and CLiDE Professional



## 2) CLiDE Batch is integrated into JChem for SharePoint and D2S

# Further information

---



**[www.keymodule.co.uk](http://www.keymodule.co.uk)**

---

**[info@keymodule.co.uk](mailto:info@keymodule.co.uk)**