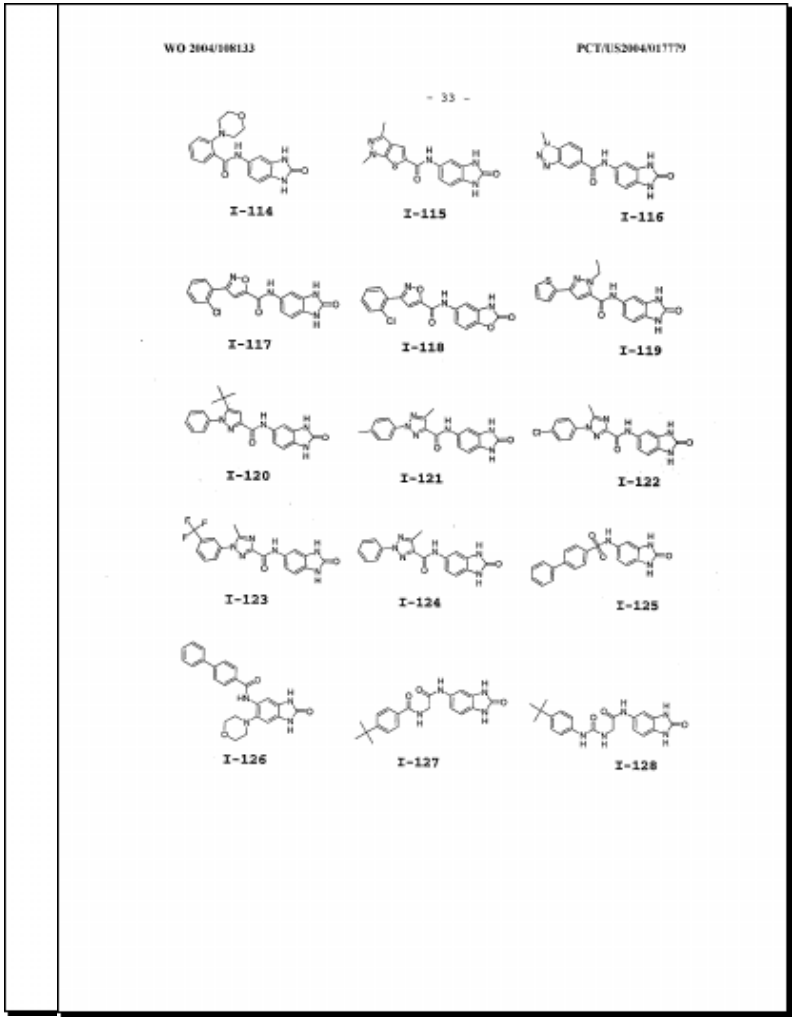
CLiDE: improved accuracy in extracting chemical structure data from documents

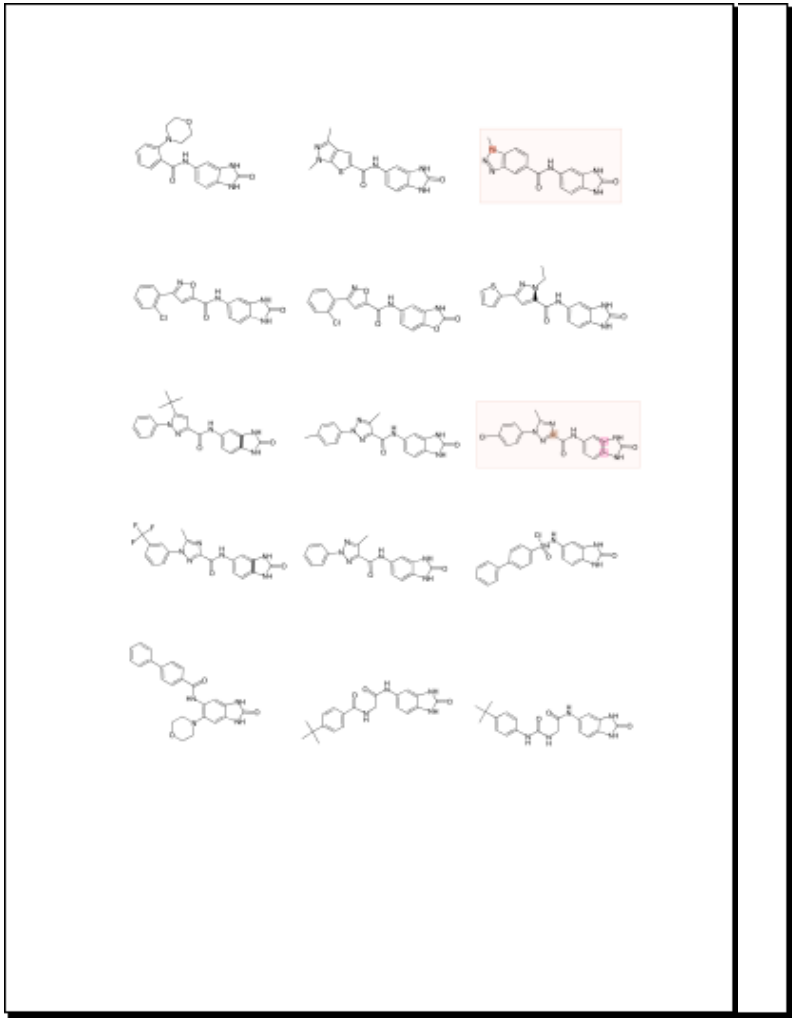
Aniko T. Valko
Keymodule Ltd.

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

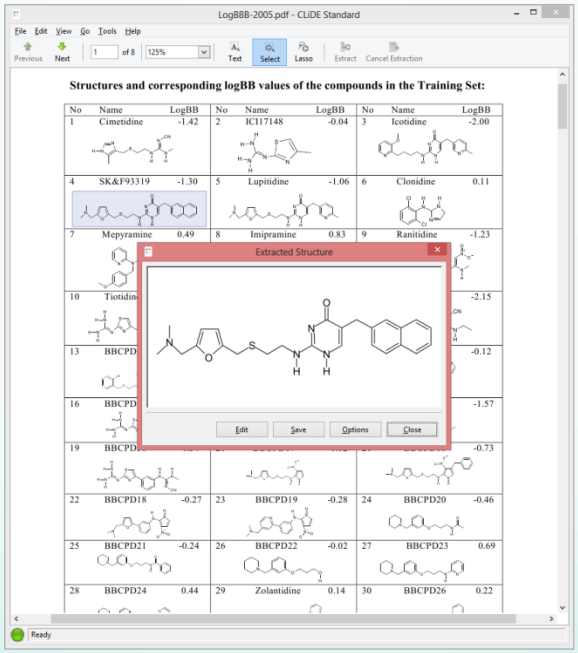




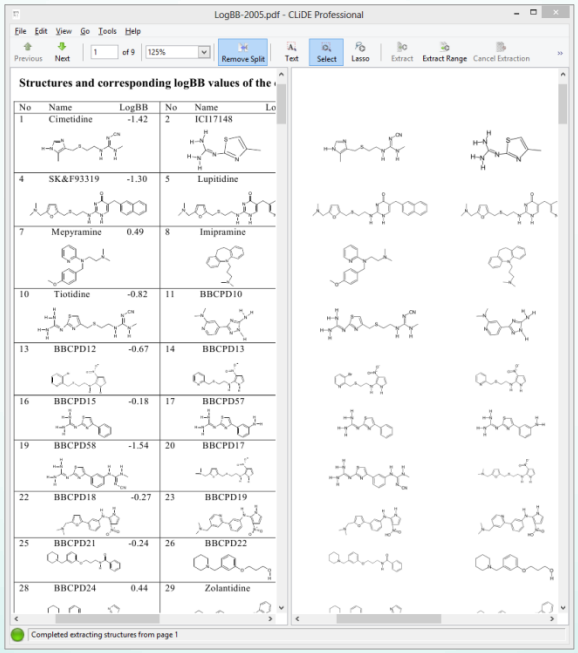
Flavours of CLiDE

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

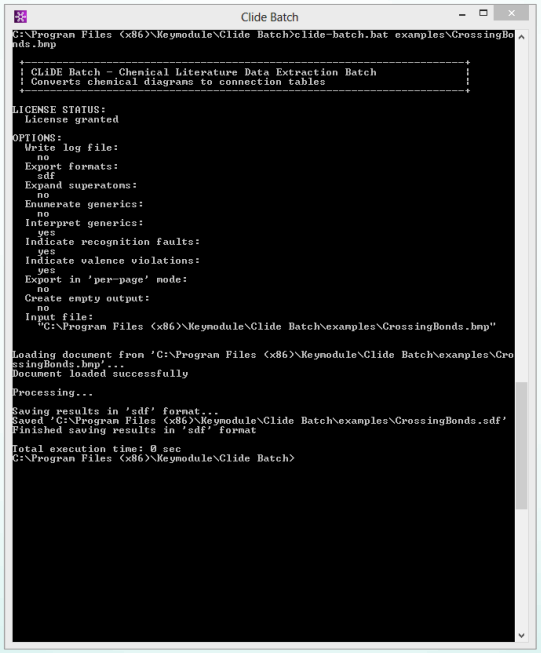
CLiDE Standard



CLiDE Professional



CLiDE Batch



Developments in CLiDE

keymodule

1991	1998	2006	present
UNIVERSITY OF LEEDS UK	SimBioSys Inc. Canada	Keymodule Ltd. 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

USPTO	**59.84%**	→	**93.81%**
Maybridge	**75.66%**	→	**90.78%**
TREC-CHEM 2011 training	**57.96%**	→	**92.84%**
TREC-CHEM 2011 topics	**57.84%**	→	**90.90%**

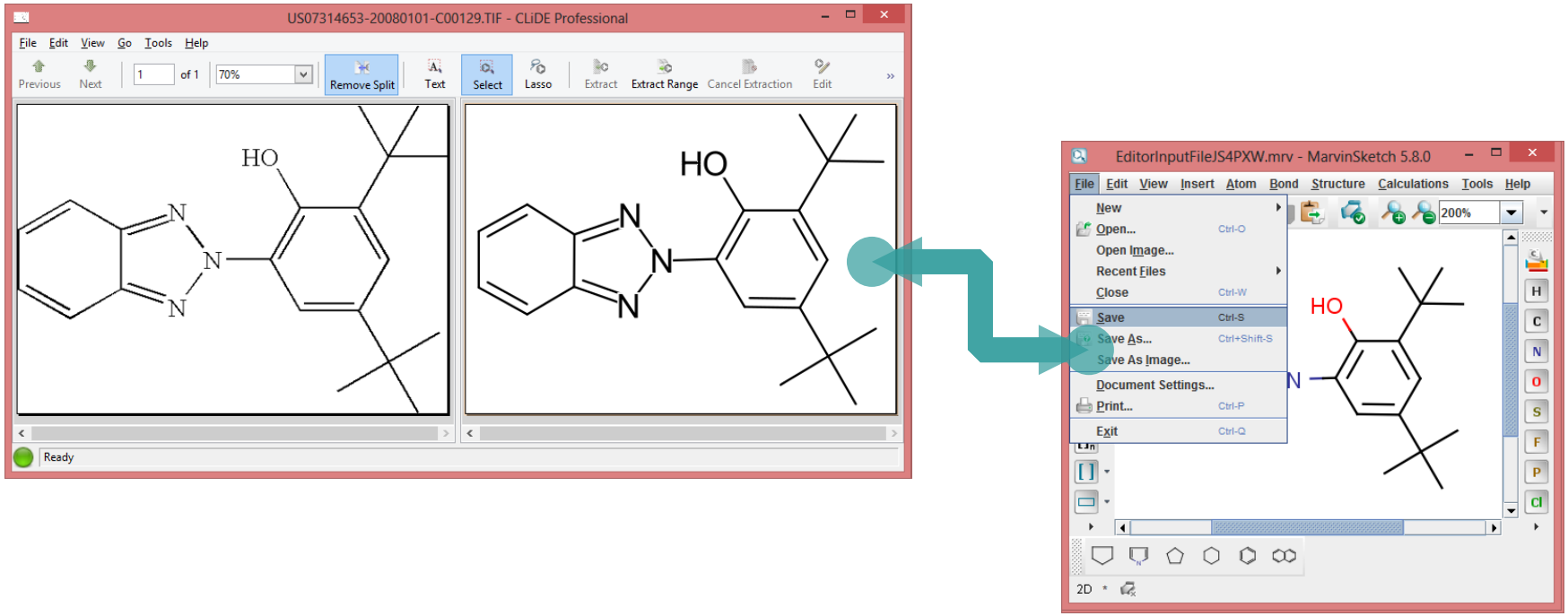
2) Non-Markush structures of two patents

WO2008099019	**12.13%**	→	**98.20%**
US6410540	**22.02%**	→	**97.71%**

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

info@keymodule.co.uk