OPEN SOURCING A WISWESSER LINE NOTATION (WLN) PARSER TO FACILITATE ELECTRONIC LAB NOTEBOOK (ELN) RECORD TRANSFER USING THE PISTOIA ALLIANCE'S UDM (UNIFIED DATA MODEL) STANDARD

Roger Sayle[1], Noel O'Boyle[1], Greg Landrum[2] and Roman Affentranger[3]

[1] NextMove Software Ltd, Cambridge, UK. [2] T5 Informatics GmbH, Basel, Switzerland.
[3] Roche Innovation Center Basel, F. Hoffmann-La Roche Ltd, Basel, Switzerland.

1. Pistoia UDM Motivation

In November 2018, the Pistoia Alliance (a non-profit consortium of pharmaceutical and life science companies) published version 5.0 of its UDM file format, an XML-based file format defined by an XSD schema. UDM (Universal Data Model) is a data exchange format that enables experimental information on chemical reactions to be shared across different systems and organizations.

One of the significant changes introduced in version 5 is that the MOLSTRUCTURE field is no longer restricted to be an MDL molfile connection table, but now allows one of five possible molecular structure formats. These are molfile (the default if no file format is specified), smiles, inchi, cdxml and wiswesser. As a proposed interchange standard, a UDM v5 file format reader needs to be able to read all of these formats in order to correctly interpret the contents of a reaction database that conforms to this specification. While open source implementations of molfile, SMILES, InChI and CDXML file format readers exist, the inclusion of Wiswesser Line Notation (WLN) as a supported/required format is potentially problematic.

2. Wiswesser Line Notation (WLN)

WLN was invented in 1949, by William J. Wiswesser, as one of the first attempts to codify chemical structure as a line notation, enabling collation on punched cards using automatic tabulating machines and early electronic computers. WLN was a forerunner to the SMILES notation used in modern cheminformatics systems, which attempted to simplify the complex rules used in WLN encoding (at the expense of brevity) to come up with an algorithmic system more suitable for implementation on computers, where historically WLN was typically encoded by hand by trained registrars.

3. WLN Syntax

WLN encoding makes use of uppercase letters, digits, spaces and punctuation.

E	Bromine atom	F	Fluorine atom
G	Chlorine atom	H	Hydrogen atom
I	Iodine atom	Q	Hydroxyl group, -OH
R	Benzene ring	S	Sulfur atom
U	Double bond	UU	Triple bond

V	Carbonyl, -C(=O)-
C	Unbranched carbon multiply bonded to non-carbon atom
K	Nitrogen atom bonded to more than three other atoms
L	First symbol of a carbocyclic ring notation
M	Imino or imido -NH- group
N	Nitrogen atom, hydrogen free, bonded to fewer than 4 atoms
O	Oxygen atom, hydrogen-free
T	First symbol of a heterocyclic ring notation
W	Non-linear dioxo group, as in $-NO_2$ or $-SO_2-$
X	Carbon attached to four atoms other than hydrogen
Y	Carbon attached to three atoms other then hydrogen
Z	Amino and amido NH_2 group
<digit>	Digits "1" to "9" denote unbranched alkyl chains
&	Sidechain terminator or, after a space, a component separator

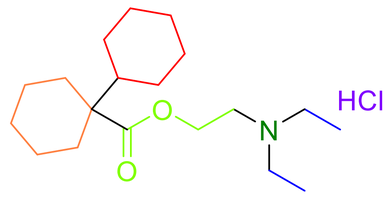
For a more complete description of the grammar, see Smith's book[2], which more accurately reflects the WLN commonly encountered than Wiswesser's book[1]. Additional WLN dialects include inorganic salts, and methyl contractions.

4. Simple WLN Examples

WLN	SMILES
4H	CCCC
1V1	CC(=O)C
WN3	[O-][N+](=O)CCC
G1UU1G	ClC#CCl
VH3	O=CCCC
NCCN	N#CC#N
ZYZUM	NC(=N)N
QY	CC(C)O
OV1 &-NA-	CC(=O)[O-].[Na+]
RM1R	c1ccccc1NCc2ccccc2
QVR BNUNR DN1&1	OC(=O)c1ccccc1N=Nc2ccc(cc2)N(C)C

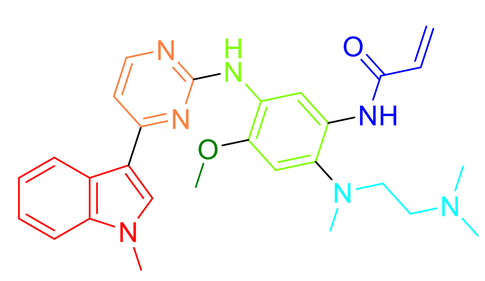
5. Advanced WLN Examples

Name: **Dicyclomine (PubChem CID 3042)**



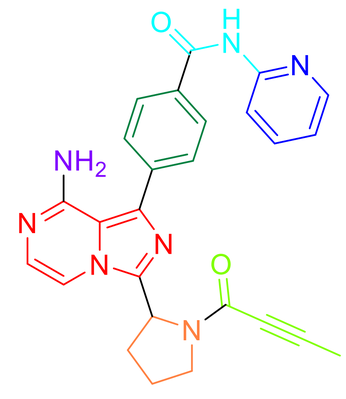
WLN: L6TJ A- AL6TJ AVO2N2&2 &GH
SMILES: CCN(CC)CCOC(=O)C1(CCCCC1)C2CCCCC2.Cl

Name: **Osimertinib**



WLN: T56 BMJ B D- DT6N CNJ BMR BO1 DN1&2N1&1 EMV1U1
SMILES: Cn1cc(c2c1cccc2)c3ccnc(n3)Nc4cc(c(cc4OC)N(C)CCN(C)C)NC(=O)C=C

Name: **(±)-Acalabrutinib**



WLN: T56 AN CN GNJ B- BT5MTJ AV1UU2& DR DVM- BT6NJ&& FZ
SMILES: CC#CC(=O)N1CCCC1c2nc(c3n2ccnc3N)c4ccc(cc4)C(=O)Nc5ccccn5

6. Benchmark Results

A C++ implementation of a WLN reader has been contributed to both RDKit and Open Babel, based on a common shared code base.

A benchmark set of 6589 WLN line-formulae was extracted from the NCBI's PubChem database by searching the depositor supplied synonyms for the substring "WLN:". The majority of these had been deposited by the NCI's DTP program. Of these, 6589 WLN over 4993 (>75.8%) can be interpreted as valid molecules and converted to SMILES or InChI.

```
% obabel -iwln -osmi -:ZVM1MVZ
NC(=O)NCNC(=O)N
% obabel -iwln -oinchi -:"1Y&Y2F1Y1QN1&1"
InChI=1S/C11H24FNO/c1-9(2)10(5-6-12)7-11(8-14)13(3)4/h9-11,14H,5-8H2,1-4H3
```

7. Future Work

- Open source a Wiswesser Line Notation writer.
- Support WLN's centrosymmetric and asymmetric multipliers.
- Support advanced (perifused, spiro and bridged) ring systems.
- Support stereochemistry via Cahn-Ingold-Prelog (CIP) descriptors[4].
- Consider tentative WLN rules: chelates, radicals, isotopes, mixtures, polymers, Markush and uncertainties (though these never/rarely appear in circulation).

8. Acknowledgements

Many thanks to Wendy Warr, Barrie Walker and Steve Heller for discussions on and memoirs of using WLN.

9. Bibliography

1. William J. Wiswesser, "**A Line-Formula Chemical Notation**", Thomas Crowell Company publishers, 1954.
2. Elbert G. Smith, "**The Wiswesser Line-Formula Chemical Notation**",McGraw-Hill Book Company publishers, 1968.
3. Stephen R. Heller and Deena A. Koniver, "**Computer Generation of Wiswesser Line Notation. II. Polyfused, Perifused and Chained Ring Systems**", *Journal of Chemical Documentation*, 12(1):55-59, 1972.
4. R. Hanson, J. Mayfield *et al.*, "**Algorithmic Analysis of Cahn-Ingold-Prelog Rules of Stereochemistry**", *J. Chem. Inf. Model.*, 58(9):1755-1765, 2018.

NEXTMOVE SOFTWARE

www.nextmovesoftware.co.uk
www.nextmovesoftware.com

NextMove Software Limited
Innovation Centre
Cambridge Science Park
Milton Road, Cambridge
England, UK CB4 0EY