MedChemica
CREATING A STEP CHANGE IN MEDICINAL CHEMISTRY

Efficiently delivering better molecule depictions & highlighting: Why, What and How.

Ed Griffen & David Cosgrove

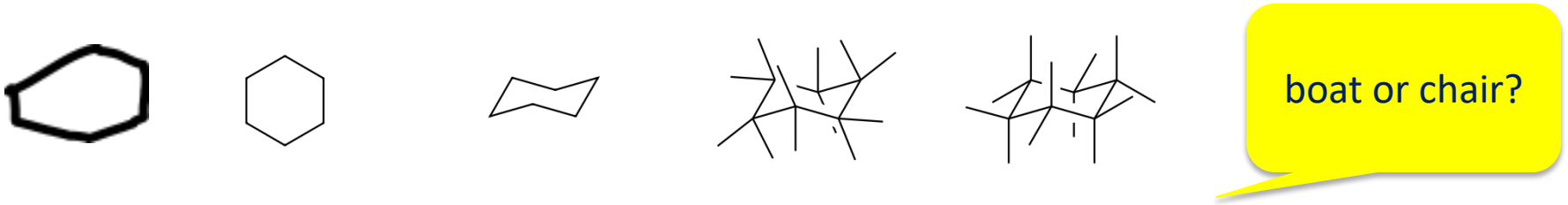
ed.griffen@medchemica.com
david@cozchemix.co.uk

@MedChemica
@covid_moonshot

www.medchemica.com

9th RDKit UGM , Virtual, 8th October 2020

Who cares about drawing molecules?

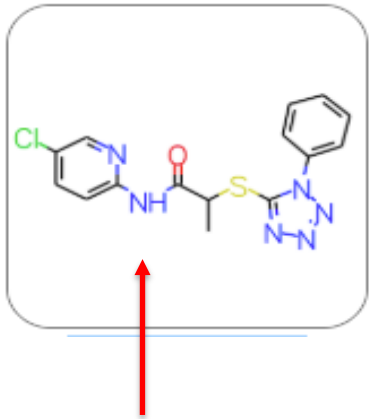


- Organic chemists judge each other on their chemical drawing – because it is not a "drawing" it is a precise chemical description.
- Software that poorly depicts molecules will be **judged as poor however good the underlying algorithms**

What we needed - as MedChemica

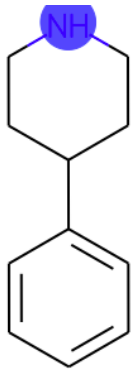
Item	Description	Priority
1	Clipping of some structures - in general smaller molecules are 'expanded' to fit the box such that edge atoms (Cl, OH examples) are clipped off.	1
2	OH NH (heteroatoms-protonated) tail groups align bond to the heteroatom (currently centred between the two atoms e.g. OH)	2
3	OH and NH (heteroatoms-protonated) - the 'wrong way' around in ring systems of Left-to-right orientation (up/down, R to L).	3
4	Define fixed bond length and font size, and only scaling down when molecule exceeds the size of a defined box. No defined box then depiction is larger.	4
5	Defined orientation of molecule - e.g. MC would like to offer the user 4 orientations of the molecule so that rendering of a powerpoint / spreadsheet yields what they want without need for editing. Consider like portrait and landscape.	5
6	Two colour highlights for atoms when smart two smarts patterns hit the same atom	6
7	Improved chiral depictions - overlap H atoms examples	7
8	SMIRK depiction - control over the bond drawing - current fixed on dotted bond only - 1 hour exploration of what this involves only	8

What these drawing crimes look like

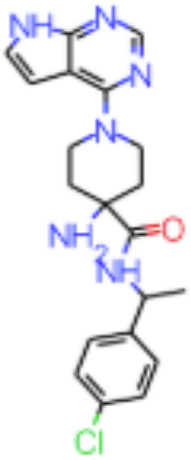


H should be below

Highlight clipped by bounding box

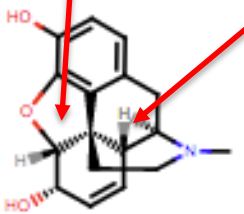


H should be above



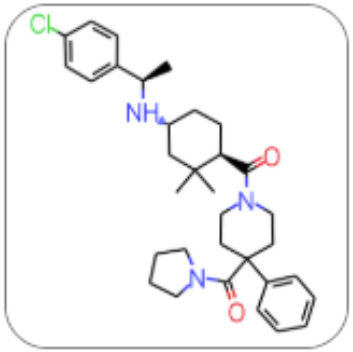
Too much chirality…

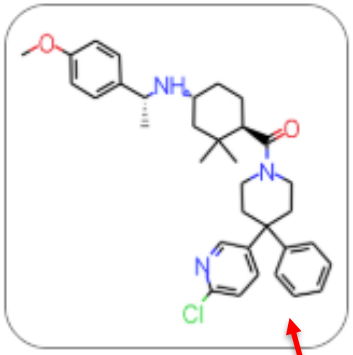
Poor H placement

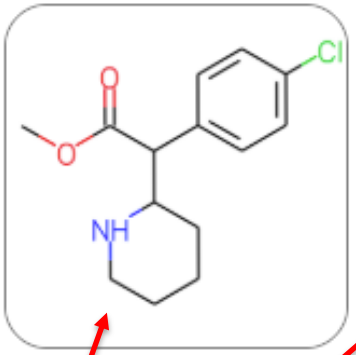


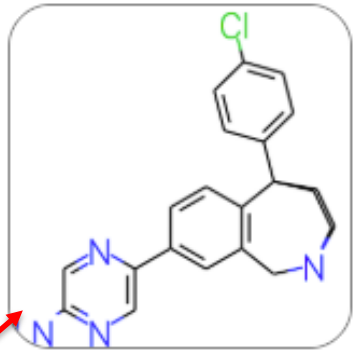
morphine

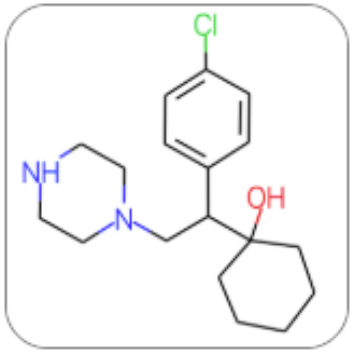
Other offences to be taken into account


CHEMBL551876
pIC50: 7.229

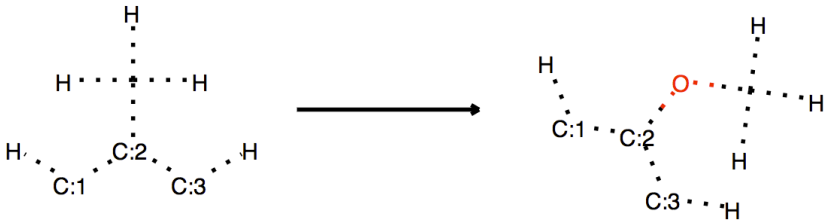

CHEMBL551206
pIC50: 8.097


CHEMBL219116
pIC50: 3.504


BDBM160604
pIC50: 8.456


CHEMBL459644
pIC50: 5.817

Consistent sizing a nightmare

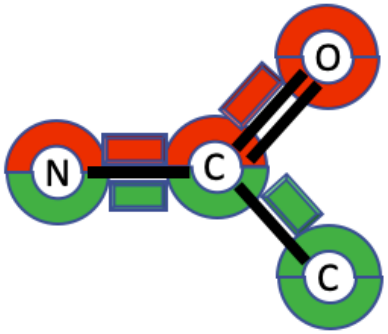


There are aliphatic and aromatic bonds in this SMIRKS…who can tell?

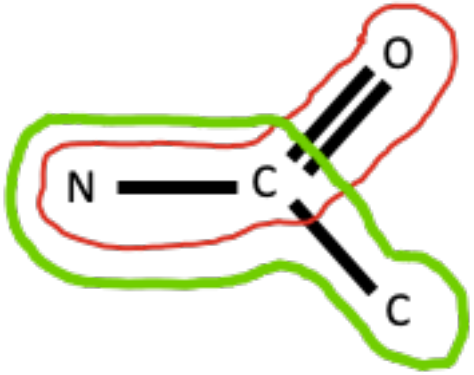
[c:1][c:2](C([H])([H])([H]))[c:3]>> [c:1][c:2](OC([H])([H])([H]))[c:3]

How to highlight multiple SMARTS matches?

- In working up SMARTS often there are multiple hits to the same structure – understanding these is critical: needs multi colours at atoms and bonds.
- Visualizing is a pain, highlighting can obscure coloured atom labels
- Early sketches (how will it deal with 3 and 4 matches on an atom though?)



Inspire by London Underground and Tokyo metro – it's a "visual grammar' people understand



Different sized lasso's also clear

What steps to take first (and why)

- Know what you want
- Talk to Greg
  - And then listen to what he says!
    - Getting the API right
- Post an Issue
  - Look at the responses
  - Adapt and build on the best ideas
  - Enables extensions

The benefits of OS software is beyond it being free – take the opportunity to crowdsource solutions early

Issues, issues

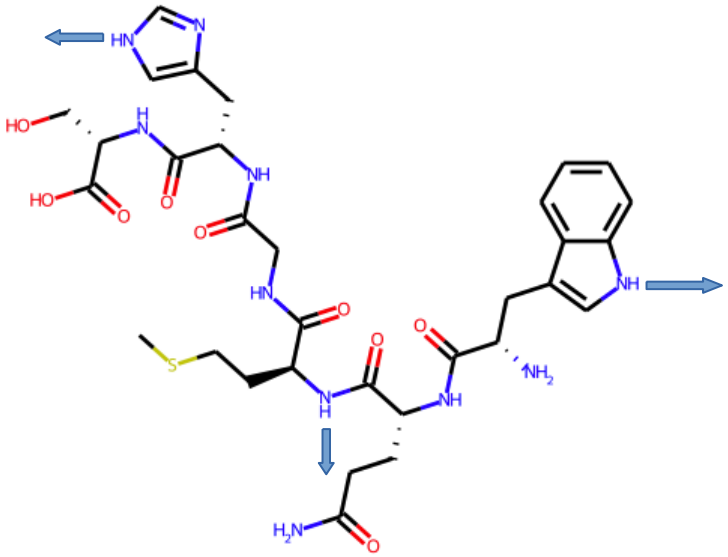
- https://github.com/rdkit/rdkit/issues/2931

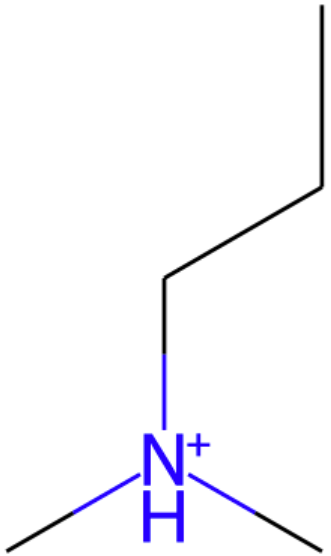
Fixing the Clipping

- Text and Highlighting not included when calculating scale
  - Scale reduced by 95%
- Calculate string rectangles and highlights as part of scale calculation
  - Requires pre-formatting of strings
  - Iterate – changing scale changes size of text and highlights
- Fix dependent on other fixes

Fixing Text

- Text centred on atom coordinates
- Code existed to say whether text should point N, S, E, W
  - Needed tweaking
- Atoms of degree 1 always E or W
- Sum vectors of attached bonds
  - Choose direction based on slope
- Atoms of degree 3 and with H



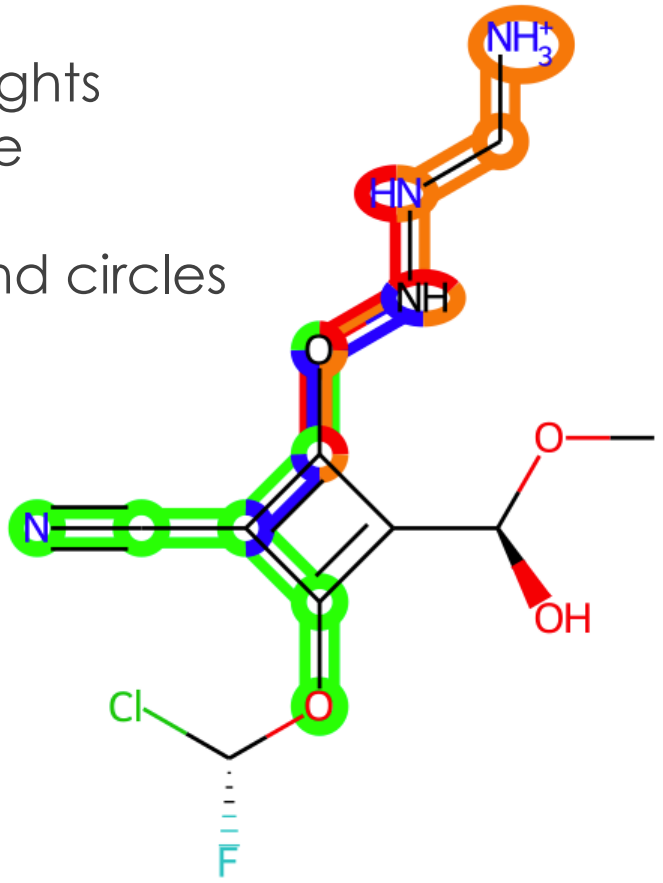


Fixing Text - 2

- Split label string into tokens
  - NH₂<super>+</super>
    - N
    - H₂<super>+<super>
- Calculate bounding rectangle of each
- Draw first token at atom coords
  - Centre on coords
- Subsequent tokens depending on orientation
- W orientation – draw backwards
- N, S orientation – left/right justify

2-Colour Highlights

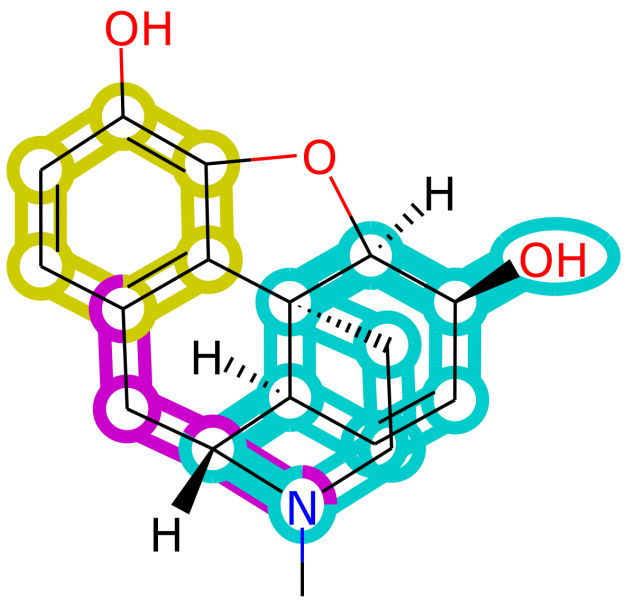
- Tube-map style
- Why stop at 2 colours?
- New API point – drawMoleculeWithHighlights
  - Rather than alter existing drawMolecule
    - don't break existing code
- Need to calculate intersection of lines and circles
  - Stackoverflow knows everything



Multi colour Highlights:– practical uses

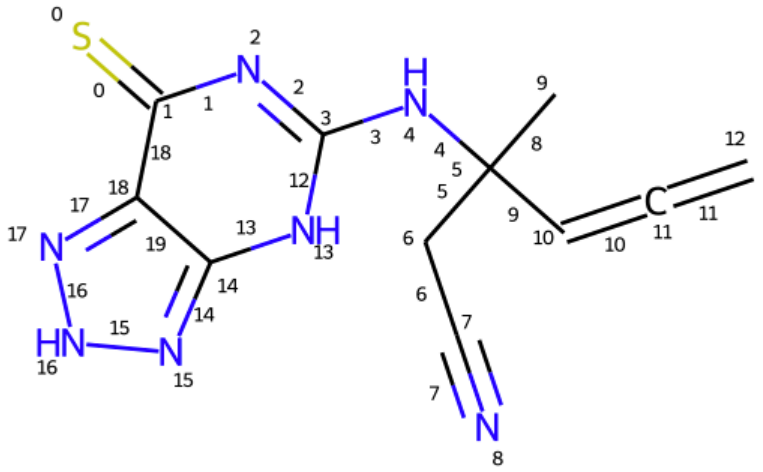
Morphine: visualizing important pharmacophores:

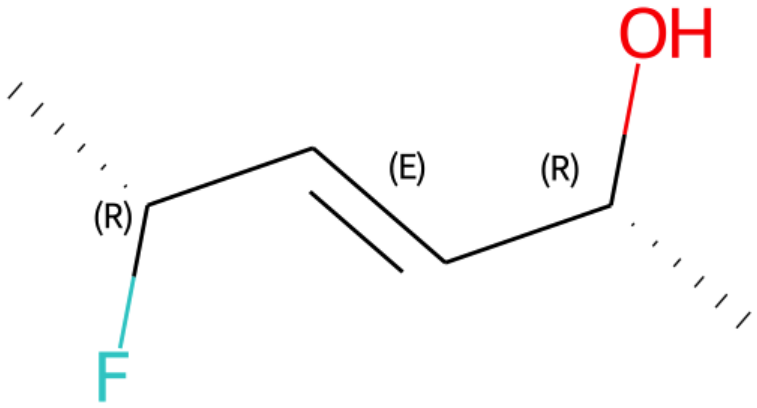
- Aryl ring **(yellow)**
- Base 3 bonds to aryl ring **(magenta)**
- Hydrogen bond donor 6 bonds to base **(cyan)**
  - *Note:* all paths round rings highlighted



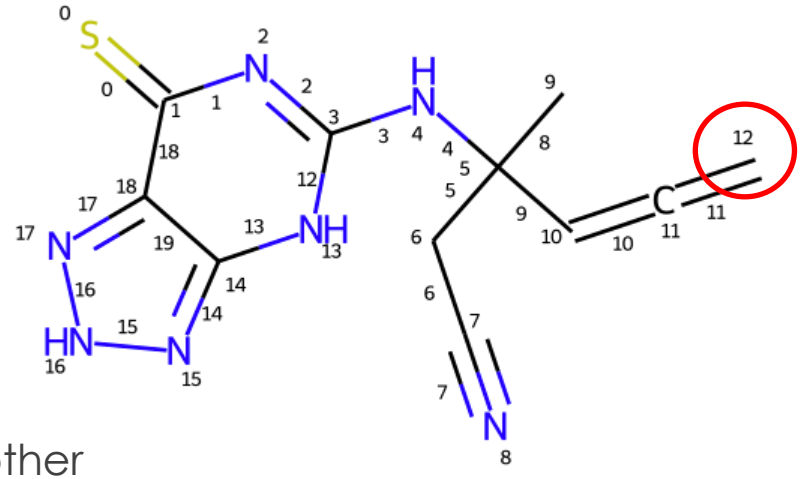
Adding Atom/Bond Annotation

- A Bonus from the 'Talk to Greg' stage
  - He funded adding annotation
- Ability to annotate with arbitrary small strings
- Canned options for
  - Atom/Bond sequence number
  - Stereo information





Adding Atom Annotation



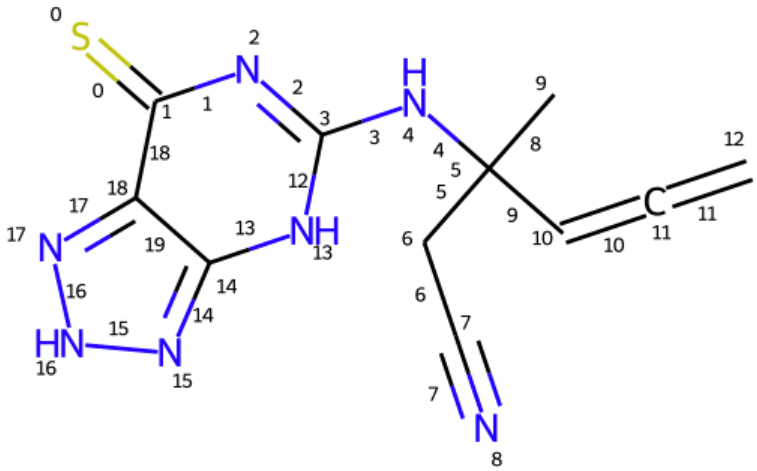
1. Pick a starting point using bonds to atom
2. See if rectangle round text clashes with bonds, other labels etc.
3. If not, use it
4. If it does, step 30° anti-clockwise
5. Goto 2
6. If returns to start point, step out 20% of bond length and goto 2

1. Increment step out twice more. If still no joy, stick it at the atom coordinates.

1. Greedy algorithm – may result in bad layout

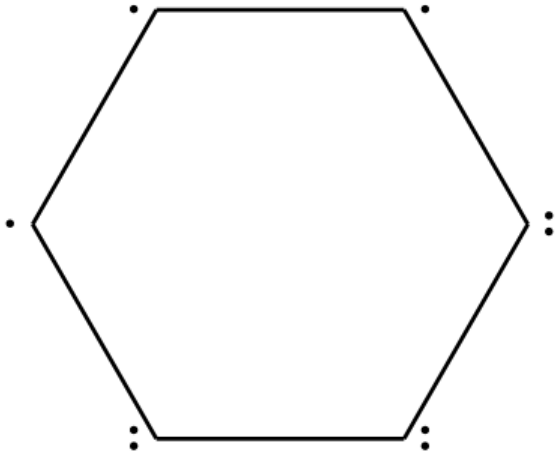
Adding Bond Annotation

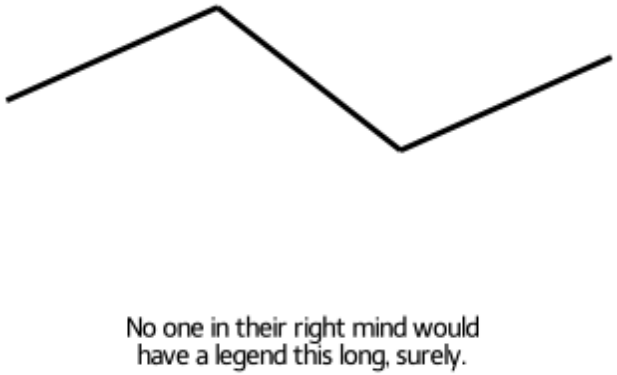
1. Start at mid-point of bond
2. Step aside perpendicular to it
3. See if bounding rectangle clashes
4. If not, use it
5. If yes, try the other side
6. If both sides fail, step out a bit and try again
7. If that fails, move towards one end
8. If that fails, move towards other end
9. If all fails, put it across the bond.



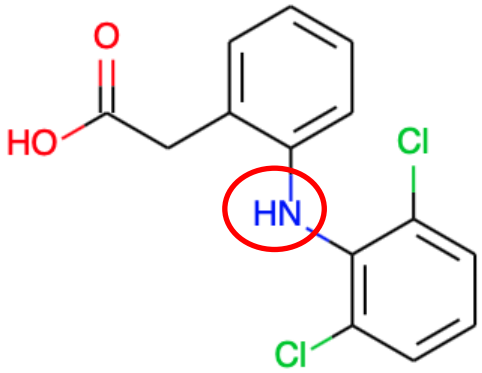
Whilst In the Code

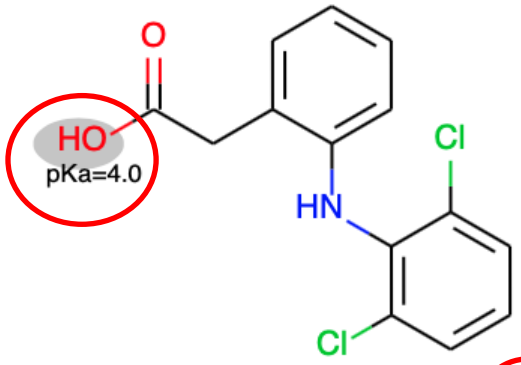
- Show radicals
- Allow multi-line titles
  - Put a new-line in the string

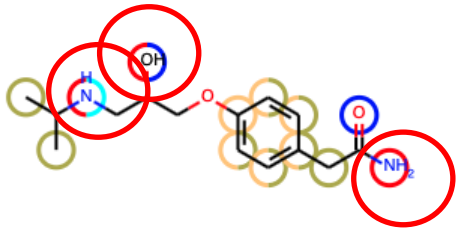


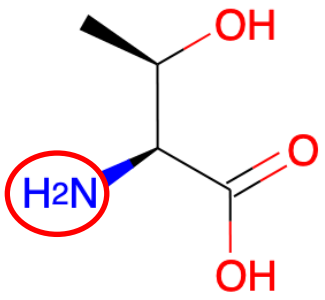


Greg's Blog Post (and a couple of others)

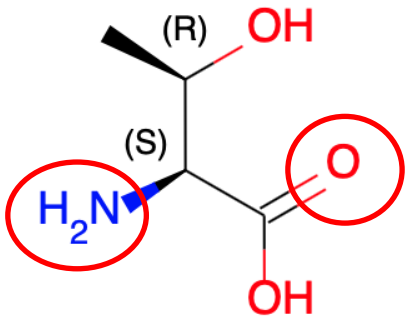




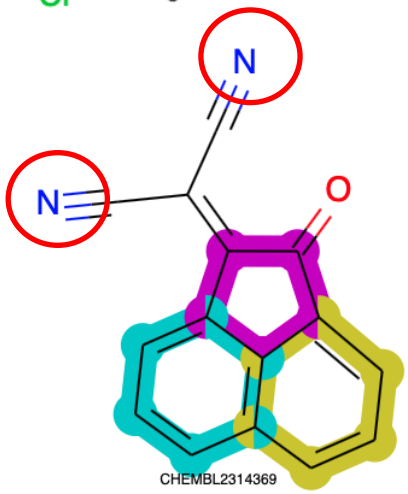




Firefox



Safari



https://rdkit.blogspot.com/2020/04/new-drawing-options-in-202003-release.html
https://iwatobipen.wordpress.com/2020/05/01/draw-molecules-as-svg-in-horizontal-layout-drawing-rdkit-memo/
https://rdkit.blogspot.com/2020/05/new-drawing-options-addendum.html

Font Metrics

- SVG and Cairo return same height bounding rectangle for each character
  - Widths vary e.g. W and I
  - Heights accommodate below line e.g. g and y
- Makes it difficult to be precise with character placement
- SVG and Cairo pictures different
- SVG in different browsers different

UK COVID Lockdown Occurs ☹

- Between Contracts
- No Holidays
- No cycling
- What does a person do…?

- Fix those "dreadful" problems in the RDKit drawings ☺

Freetype (freetype.org)

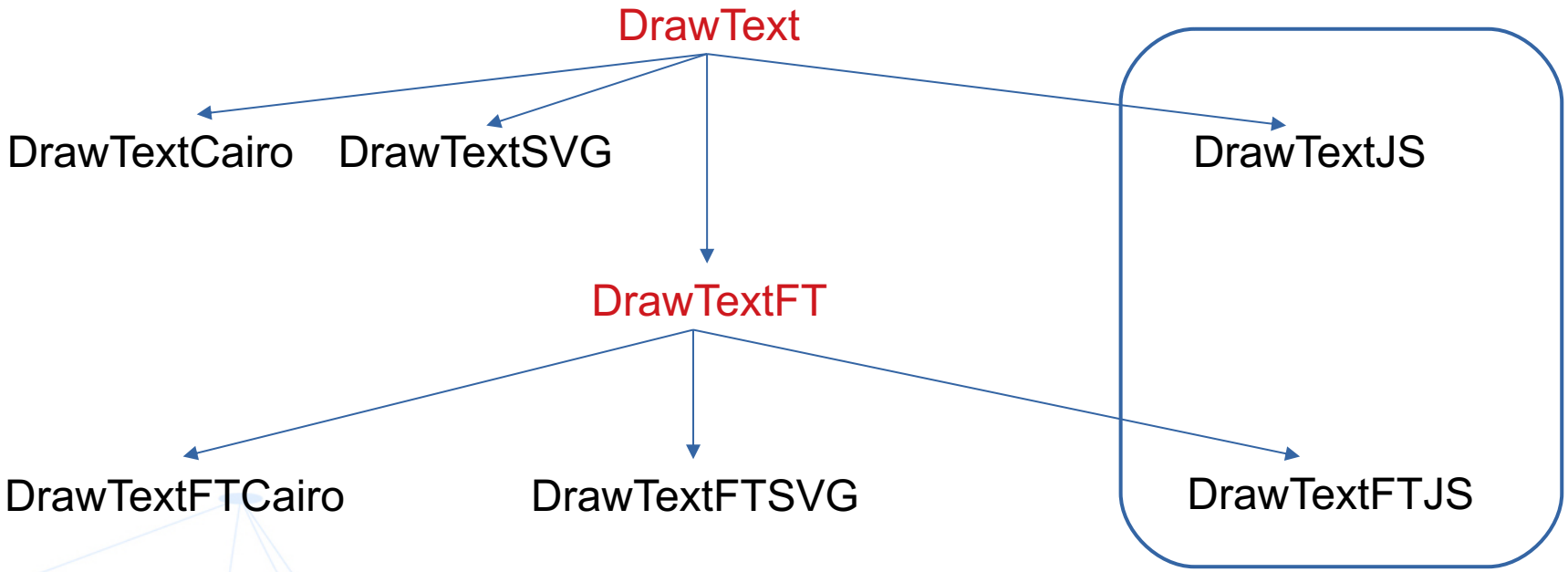
- "Freely available software library to render fonts"
- Versions for Mac, Linux and Windows10
- Several font formats (TrueType, CFF, WOFF 2…)
- Size information for each character
- C API

open-foundry.com

- Huge number of open-source fonts to download

Implementing FreeType - 0

- New classes for drawing text into MolDraw2D object



Greg

Implementing FreeType - 1

- FreeType draws character (glyph) using supplied callbacks:
  - moveToFunction
    - Place drawing cursor to position
  - lineToFunction
    - Draw line from drawing cursor to new position
  - conicToFunction
    - Draw quadratic Bezier function between drawing cursor and new position
  - cubicToFunction
    - Draw cubic Bezier function between drawing cursor and new position

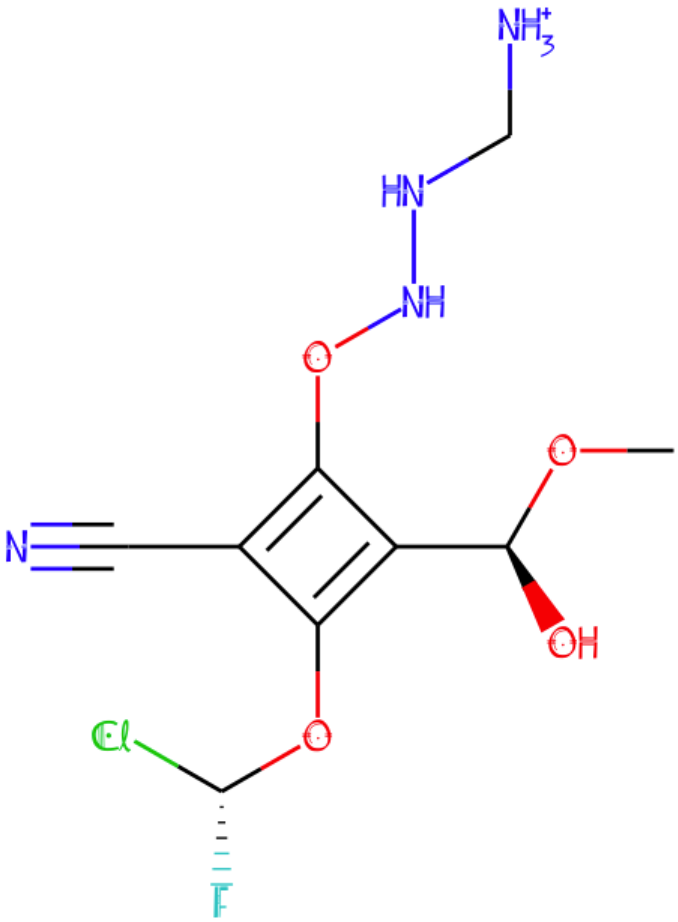
Implementing FreeType - 2

- Abstract Base Class DrawTextFT
  - Loads font
  - Calculates rectangles round glyphs
  - Draws strings using FreeType callbacks
  - Default font (Telex) hard-coded in
- Concrete Classes DrawTextFTCairo etc
  - Implement 4 callback functions in their API
- FreeType is C API, RDKit is C++
  - can't use class member functions directly
  - Smuggle 'this' in as user data
  - Free function callback extracts 'this' and uses to call class function

Which Way to Positive Y?

- Long-standing convention that origin is top left, positive Y goes down
  - From the old days of CRTs?
- FreeType eschews this convention
- The glyph must be inverted in FreeType coords or it comes out upside down
- Translate glyph down for subscript in FreeType coords, bounding rectangle goes the other way in draw coords

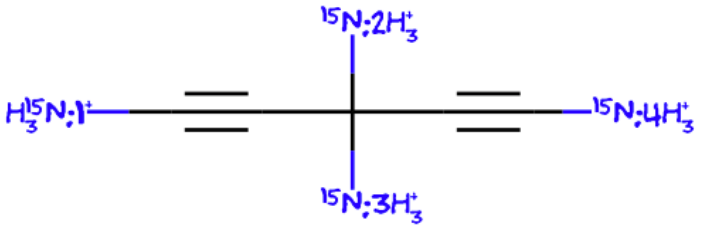
Rock Your Own Font Choice



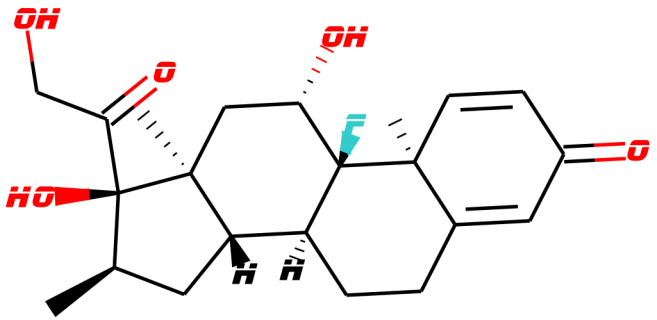
The Amadeus Font – available in the RDKit

Rock Your Own Font Choice - 2

```
smiles = '[15NH3+:1]-C#C-C([15NH3+:2])([15NH3+:3])-C#C-[15NH3+:4]'
mol = AllChem.MolFromSmiles(smiles)
drawer = rdMolDraw2D.MolDraw2DSVG(400, 200)
do = rdMolDraw2D.MolDrawOptions()
do.fontFile = 'Humor-Sans-1.0.ttf'
drawer.SetDrawOptions(do)
rdMolDraw2D.PrepareAndDrawMolecule(drawer, mol)
drawer.FinishDrawing()
with open('humor_sans.svg', 'w') as f:
    f.write(drawer.GetDrawingText())
```

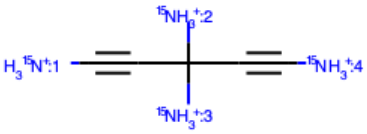


Or something more ~~2049~~ *2020*:

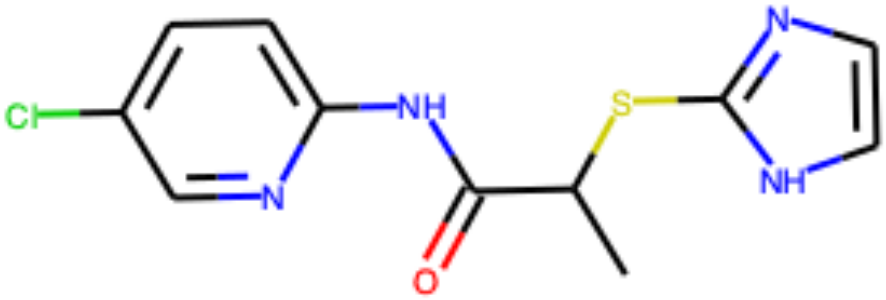


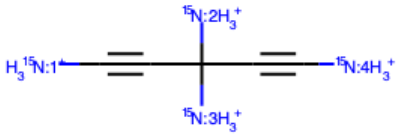
DEXAMETHASONE

So Where Are We?

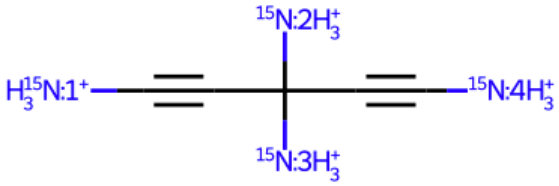


2019.09

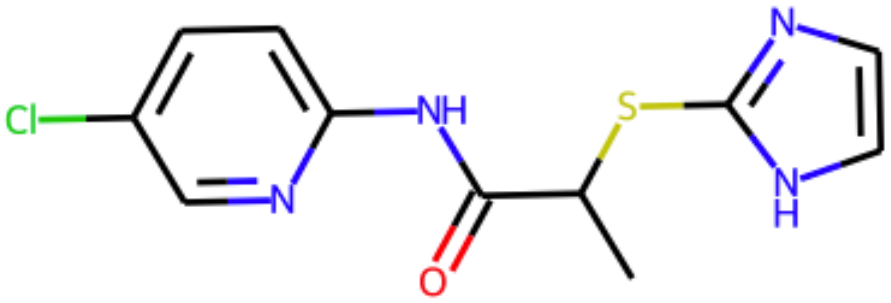




2020.09



2020.09



What we learnt

- Positives:
  - Talk to Greg early
  - Post an issue
  - Being focussed means you can have changes done efficiently
  - RDKit UGM 2019 talked to Greg → live in 2020.03 release

- To do better next time
  - Test Early
  - Test Often
  - Don't accept any more extensions till the basics are fully tested…
  - Double your initial time estimate, at least

Will MedChemica do this again?

- Absolutely
- Efficient use of time and money to make improvements that we need in non IP sensitive code base.
- Honour bound: – we use the RDKit → we contribute to its extension

Acknowledgements

- Greg Landrum
- Brian Kelley
- Al Dossetter
- All the complaining chemists….