

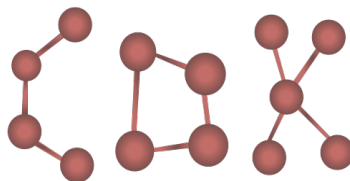


HIGHER QUALITY CHEMICAL DEPICTIONS:

Lessons Learned and Advice

John Mayfield
NextMove Software Ltd





Chemistry **D**evelopment **K**it

- Java Library, KNIME Nodes, RCDK
- **16** years old
- **115** contributors
- History in Computer Assisted Structure Elucidation

Many **legacy APIs** and bad wrong **algorithms** and **data structures**.
Many of my contributions have focussed on core functionality because I needed it for my PhD at the time

*“Every **PhD student** in cheminformatics writes their own toolkit”*

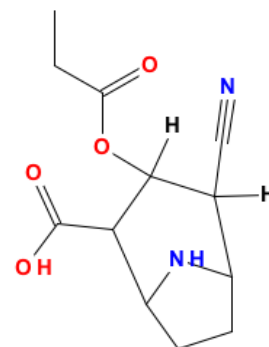
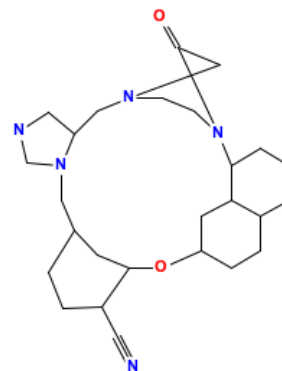
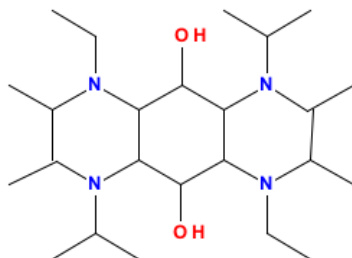
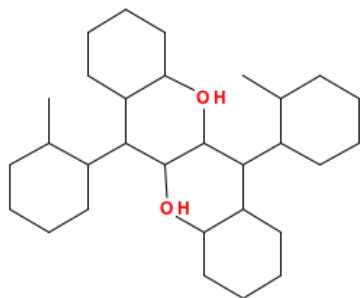
*“Every **company** in cheminformatics writes their own toolkit”*

During writing of thesis (**2013**) I needed **publication quality** depictions.
Existing **FOSS** and affordable **commercial** offerings below par (didn't want a ChemDraw license for such a short period)

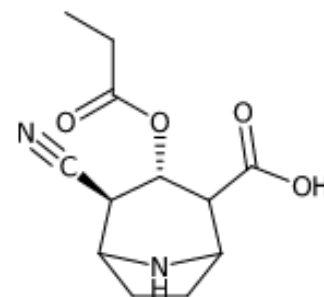
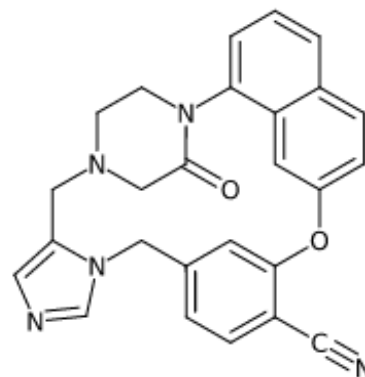
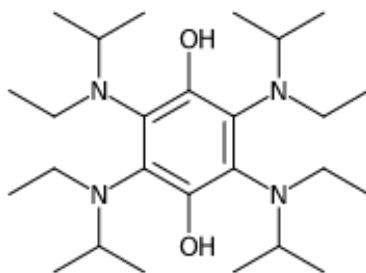
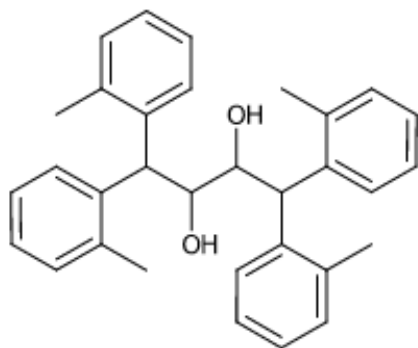


Improvements

1.4



2.0



“John will show us what good coordinate generation looks like” - Greg Landrum

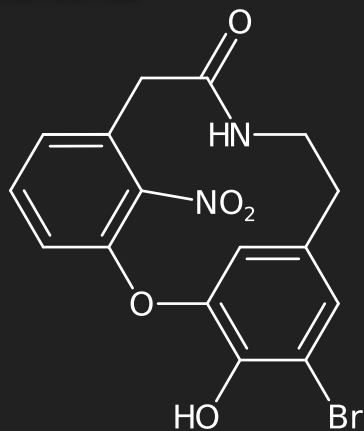
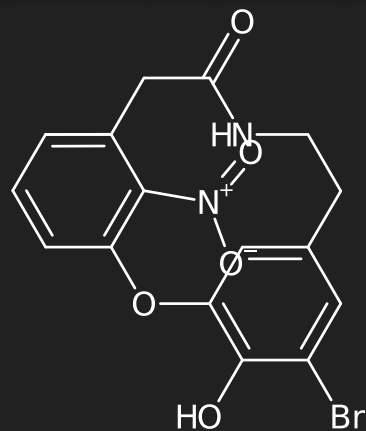


LAYOUT

"Structure Diagram Generation"

- Position **atoms** X,Y coords
- Orientation
- Wedge assignment
- Objective (overlaps)
- Subject (orientation)

OPEN PROBLEM:



RENDERING

"Drawing"

- Generate and position **graphic primitives**
- Atom Label alignment
- Fonts
- Annotation coordinates
- Display Shortcuts (Abbreviations)
- Subjective (color, donuts)



Homework

2D Layout Literature

Clark, A *et al.* **2D Structure Depiction.** *J. Chem. Inf. Model.* 2006. 46(3)

Helson, H. **Structure Diagram Generation.** *Reviews in Computational Chemistry, Volume 13.* 1999. Ch 6

Weininger, D. **SMILES. 3. Depict. Graphical Depiction of Chemical Structures.** *J. Chem. Inf. Comput. Sci.* 1990. 30(3).

Rendering Literature

Brecher J. **Graphical Representation Standards For Chemical Structure Diagrams (IUPAC Recommendations 2008).** *Pure Appl. Chem.* 2008. 80(2)

Clark, A *et al.* **Rendering Molecular Sketches for Publication Quality Output.** *Molecular Informatics.* 2013. 32

Cambridge Soft. **CDX File Format.** *Online:* <http://www.cambridgesoft.com/>

Clark, A *et al.* **Basic primitives for molecular diagram sketching.** *J. Cheminf.* 2010. 2(8)

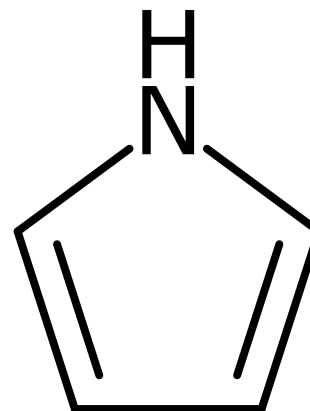
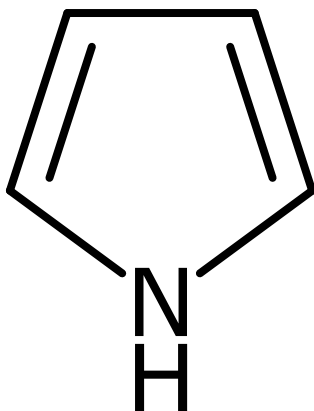
Gushurst, A *et al.* **The Substance Module: The Representation, Storage, and Searching of Complex Structures.** *J. Chem. Inf. Comput. Sci.* 1991. 31.



LAYOUT

RDKit algorithm, better architecture than **CDK**. My recent patches treat “symptom over cause”, but useful:

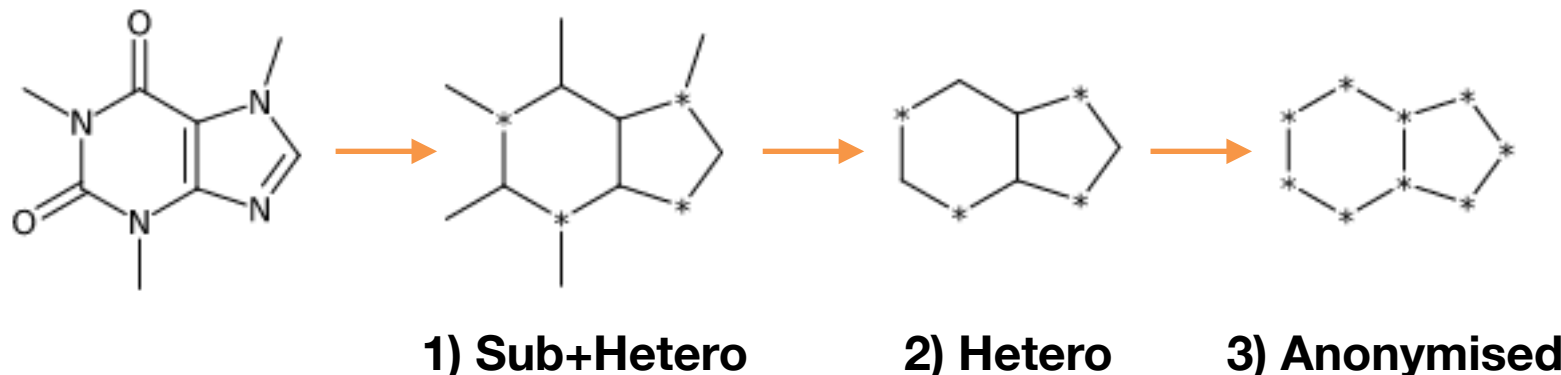
- (1) Ring templates
- (2) Macrocyclic templates
- (3) Layout refinement (fix collisions)
- (4) Humpty Dumpty



WHICH WAY UP?



Canonical Ring Indexing



Each template ring system is **indexed** in three ways,
lookup follows the same order

Capture standard orientations (algorithm fallback possible)

Generated library from hand drawn structures - duplication
needed (algorithm or hand curated)

Possible sources: **ChEBI, Suppliers, Patents, Journals**

Stored as CXSMILES in CDK

```
*1**C*1 | (1.21, .39, ; .75, -1.03, ; -.75, -1.03, ; -1.21, .39, ; .0, 1.28, ) |
```

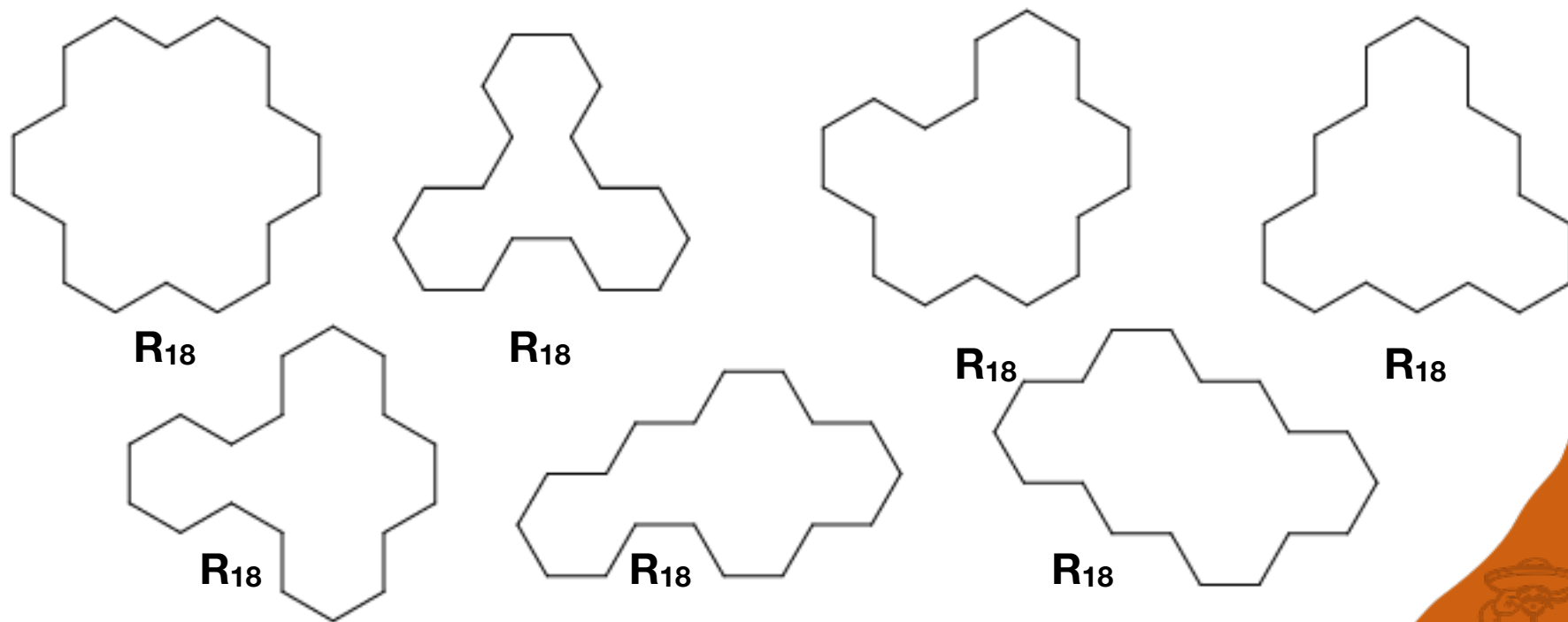


Macrocycle Indexing

Index **multiple** layouts for **even** ring size

Selects **multiple** templates and scores **registry shifts** based on: ring attach points, cis/trans correctness, heteroatom positions.

Odd size rings use the **$n+1$** template, last coord dropped.



RDDepictor.cpp

```
computeInitialCoords(mol, coordMap, efrags);

std::list<EmbeddedFrag>::iterator eri;
// perform random sampling here to improve the density
for (eri = efrags.begin(); eri != efrags.end(); eri++) {
    // either sample the 2D space by randomly flipping rotatable
    // bonds in the structure or flip only bonds along the shortest
    // path between colliding atoms - don't do both
    if ((nSamples > 0) && (nFlipsPerSample > 0)) {
        eri->randomSampleFlipsAndPermutations(
            nFlipsPerSample, nSamples, sampleSeed, 0, 0.0, permuteDeg4Nodes)
    } else {
        eri->removeCollisionsBondFlip();
    }
}

for (eri = efrags.begin(); eri != efrags.end(); eri++) {
    // if there are any remaining collisions
    eri->removeCollisionsOpenAngles();
    eri->removeCollisionsShortenBonds();
}
```

Layout Refinement

RDKit

1. Initialise
2. Sample or Rotate
3. Shrink and Bend
4. Orientation

CDK

1. Initialise
2. Rotate, Bend, Stretch, Invert
3. Orientation

Rotate: flip rotatable bonds (most desirable, optimal)

Bend: unsnap/open angles

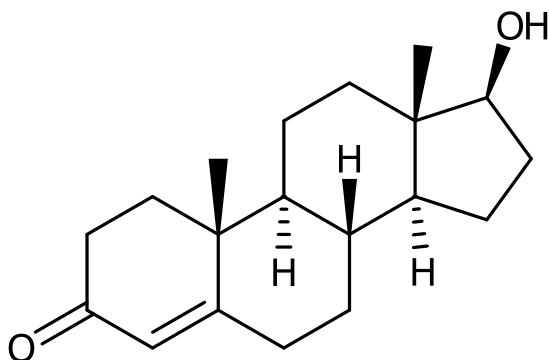
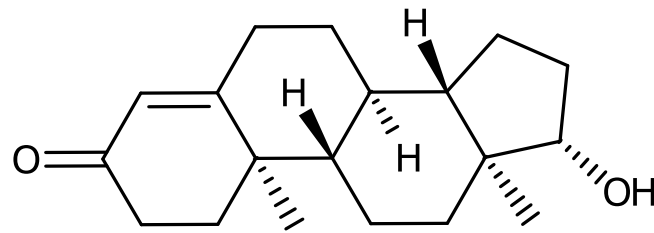
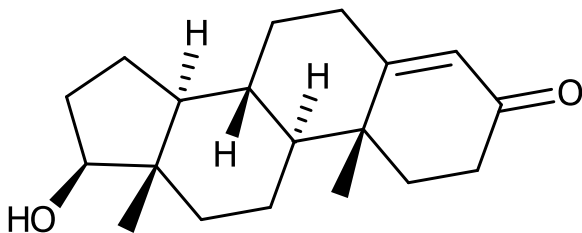
Stretch: make bonds longer

Shrink: make bonds shorter

Invert: mirror a terminal bond inside a ring



Orientation

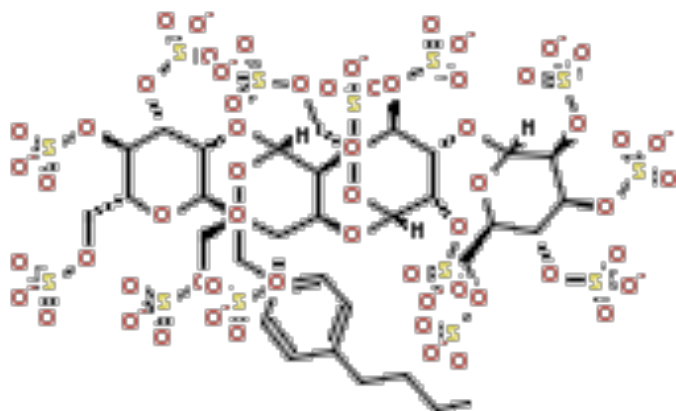


Mostly align principle ring system, rules ~IUPAC naming

- Core ring orientation (fused rings, steroids)
- Layout width/height (RDKit canonical orientation)
- Bond snapping (align to 30°)
- Symmetry (*patented in US*)



Humpty Dumpty sat on a wall,
Humpty Dumpty had a great fall.
All the king's horses and all the king's men
Couldn't put Humpty together again.



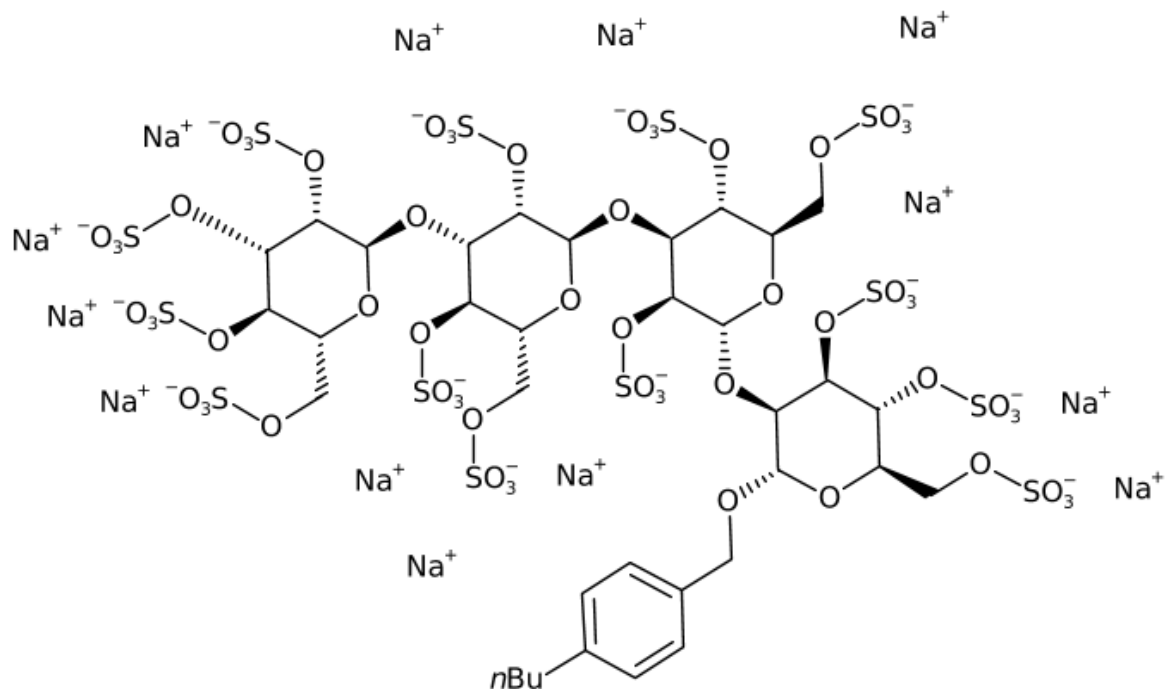
Na⁺ Na⁺ Na⁺ Na⁺ Na⁺ Na⁺ Na⁺ Na⁺ Na⁺ Na⁺ Na⁺ Na⁺ Na⁺ Na⁺

CHEMBL590010
(BioVia)

Rebond **monoatomic** ions before initial layout, delete after.



Humpty Dumpty sat on a wall,
Humpty Dumpty had a great fall.
All the king's horses and all the king's men
Couldn't put Humpty together again.



CHEMBL590010
(CDK)



Layout Comparison

Open Babel

Avalon

RDKit

CDK

Indigo

Structure Layout Testset

Based on **28** structures the **10** obstacles from Clark, A. *et al.* 2006

- 1 **new** obstacle (**11** total)

- 20 **new** structures (**48** total) - have a lot more

Previous post by **Noel O'Boyle** used random PubChem sample, too easy: <http://baoilleach.blogspot.co.uk/2008/10/cheminformatics-toolkit-face-off.html>

All layout algorithms make **mistakes** and produce **crowded** or even **misrepresent** (wrong) structures! CDK definitely still does this but also commercial offerings (see Clark 06). Emphasis here is on **silly mistakes** rather than **perfect layout**.

Evaluating **layout only**, all **rendered** with CDK here



48 Structures 11 Obstacles

1. Find Optimal Solution - avoidable overlaps **(+2)**
2. Suboptimal Solution - unavoidable overlaps **(+7)**
3. Global Chain Blocks
4. Double Bond Stereochemistry **(+3)**
5. Congested Small Rings
6. Bonding Counterions **(+2) *new***
7. Spirocenters
8. Macrocycles **(+3)**
9. Ring Template Matches **(+1)**
10. Planar Embedding
11. 3D Ring Systems **(+2)**



Performance

All **48** Structures

“Fair” **41** Subset

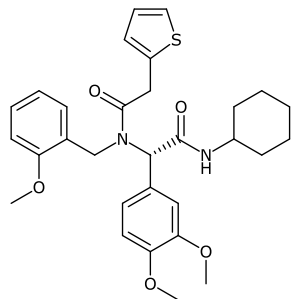
(skip 3D rings)
(skip dendritic structure)

Library	Elapsed	Mean	Elapsed	Mean
Open Babel v2.4.1	25:59.0	-	1.9s	46ms
RDKit V2016.09.1.dev1	10.2s	214ms	0.1s	2ms
Avalon 1.2.0	0.3s	6ms	0.06s	1ms
CDK 2.0-SNAPSHOT	0.5s	9ms	0.06s	1ms
Indigo 1.2.3.r0 no-smart-layout	1.7s	35ms	0.05s	1ms

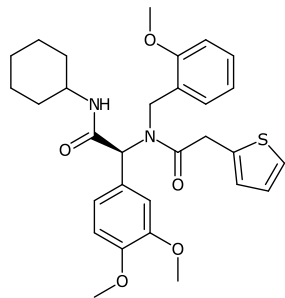
Indigo ‘smart-layout’ option: better macrocycles but a lot worst in general and degrades performance

Level 1 - Find Optimal Solution

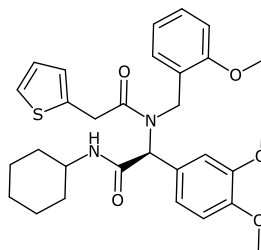
Open Babel



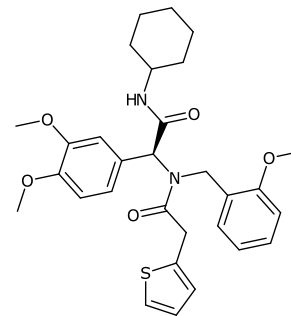
Avalon



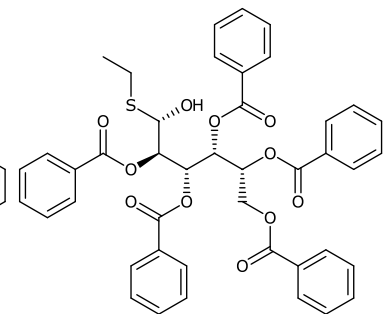
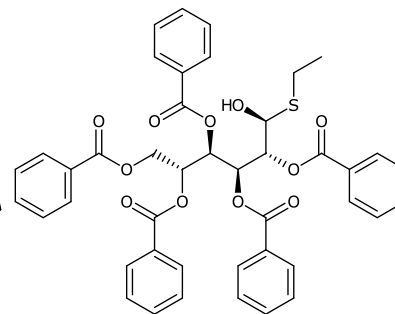
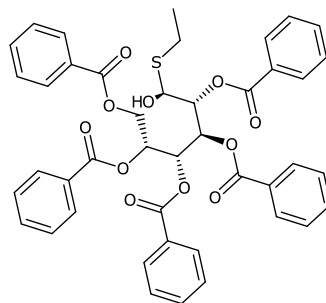
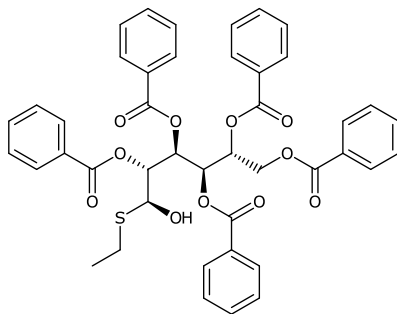
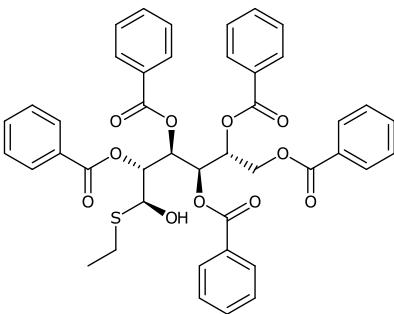
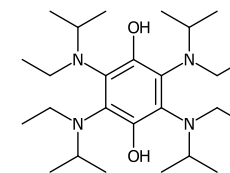
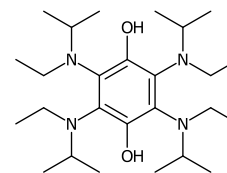
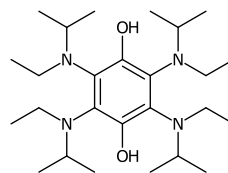
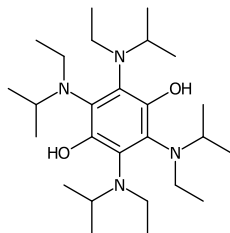
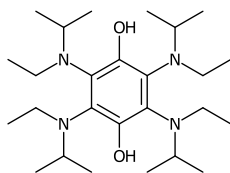
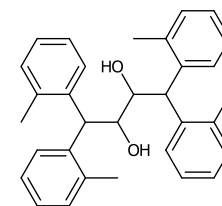
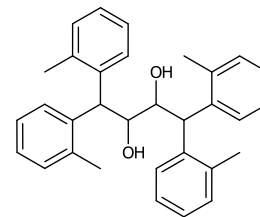
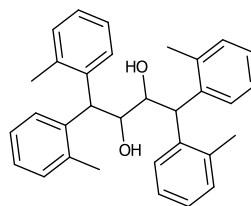
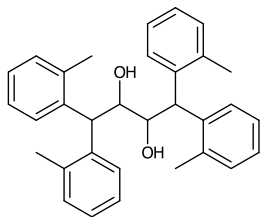
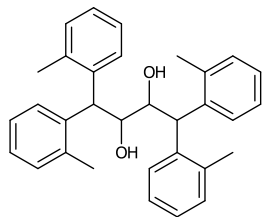
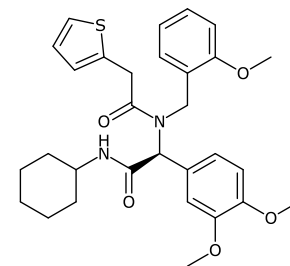
RDKit



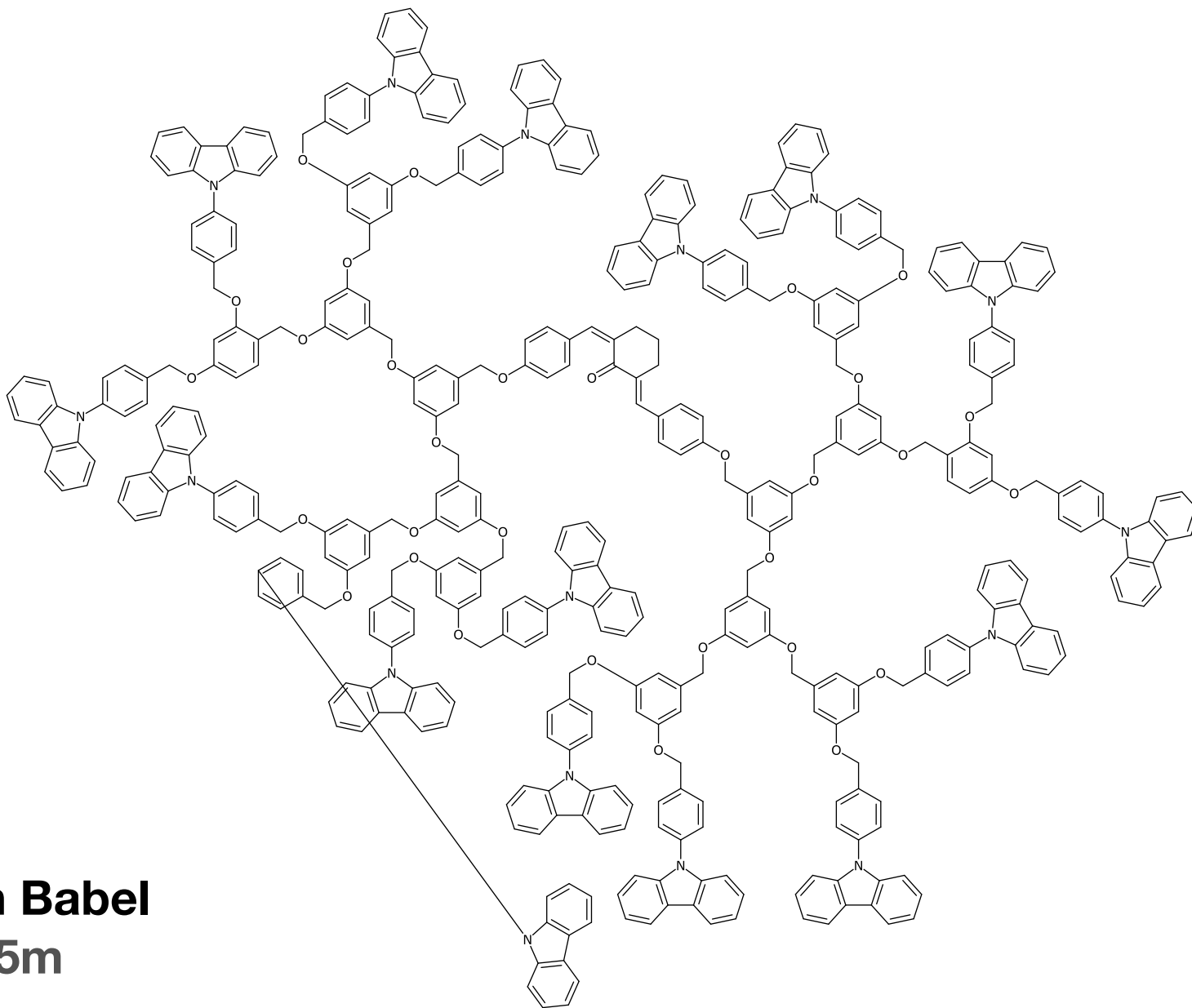
CDK



Indigo

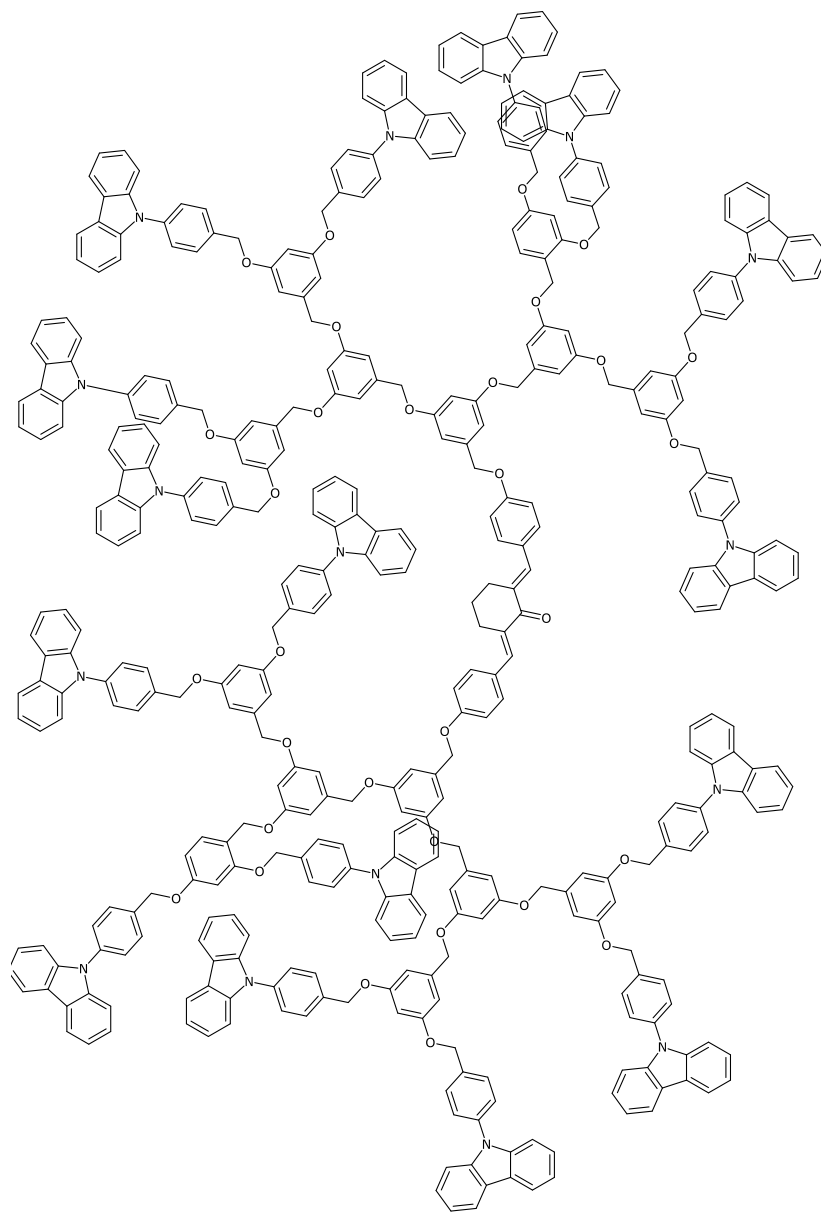


CHEMBL1077020



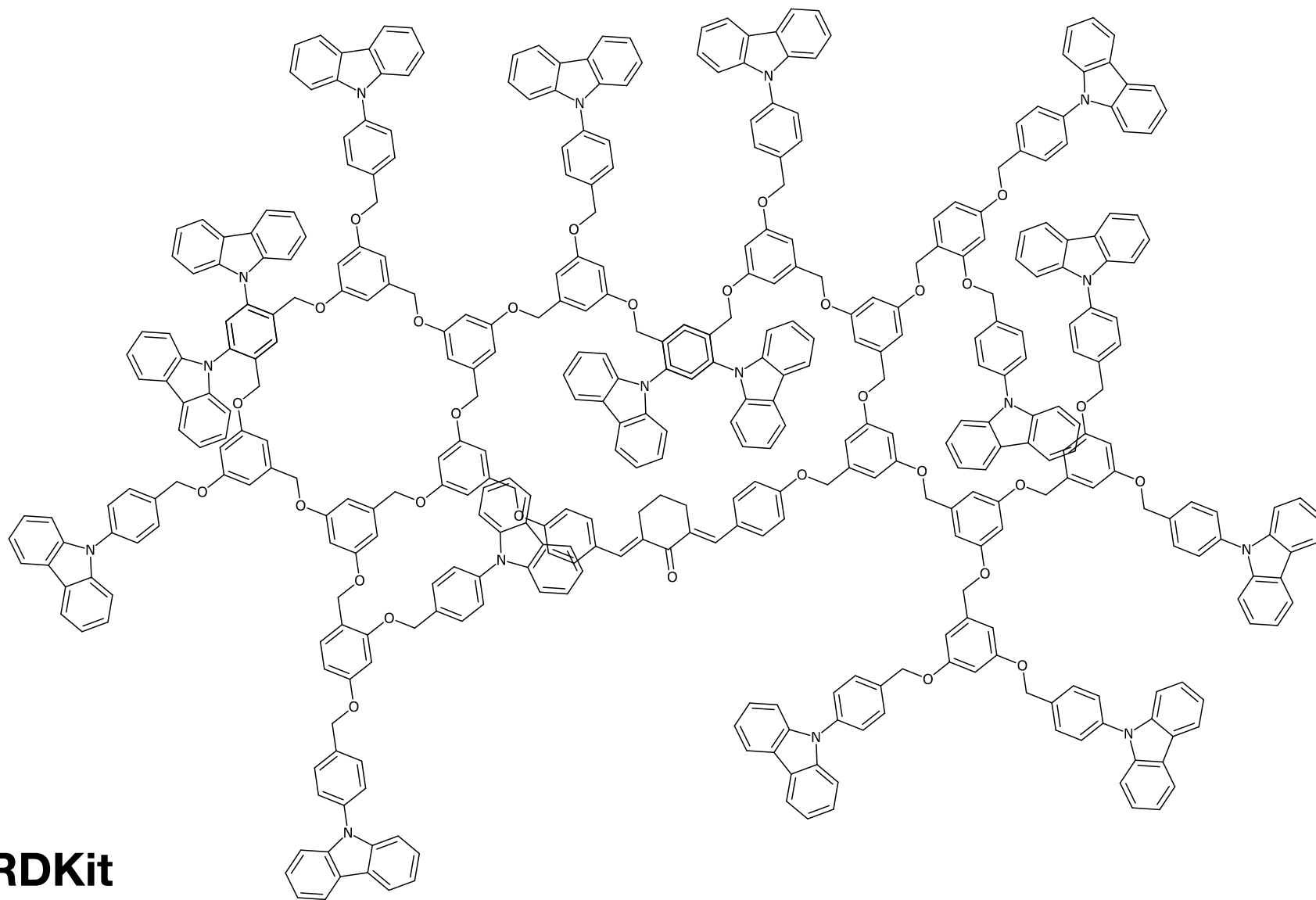
Open Babel
25m

CHEMBL1077020



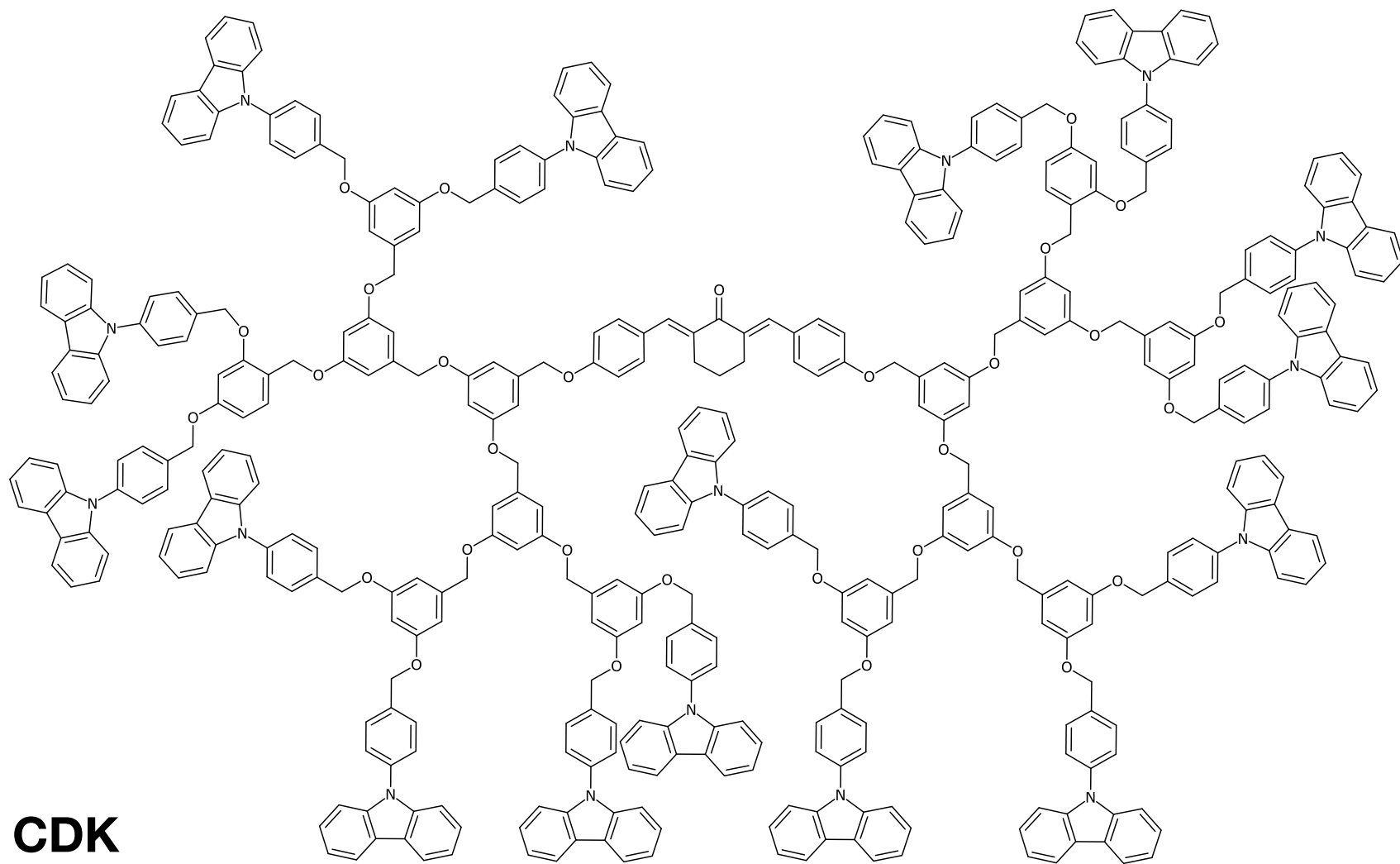
Avalon
200ms

CHEMBL1077020



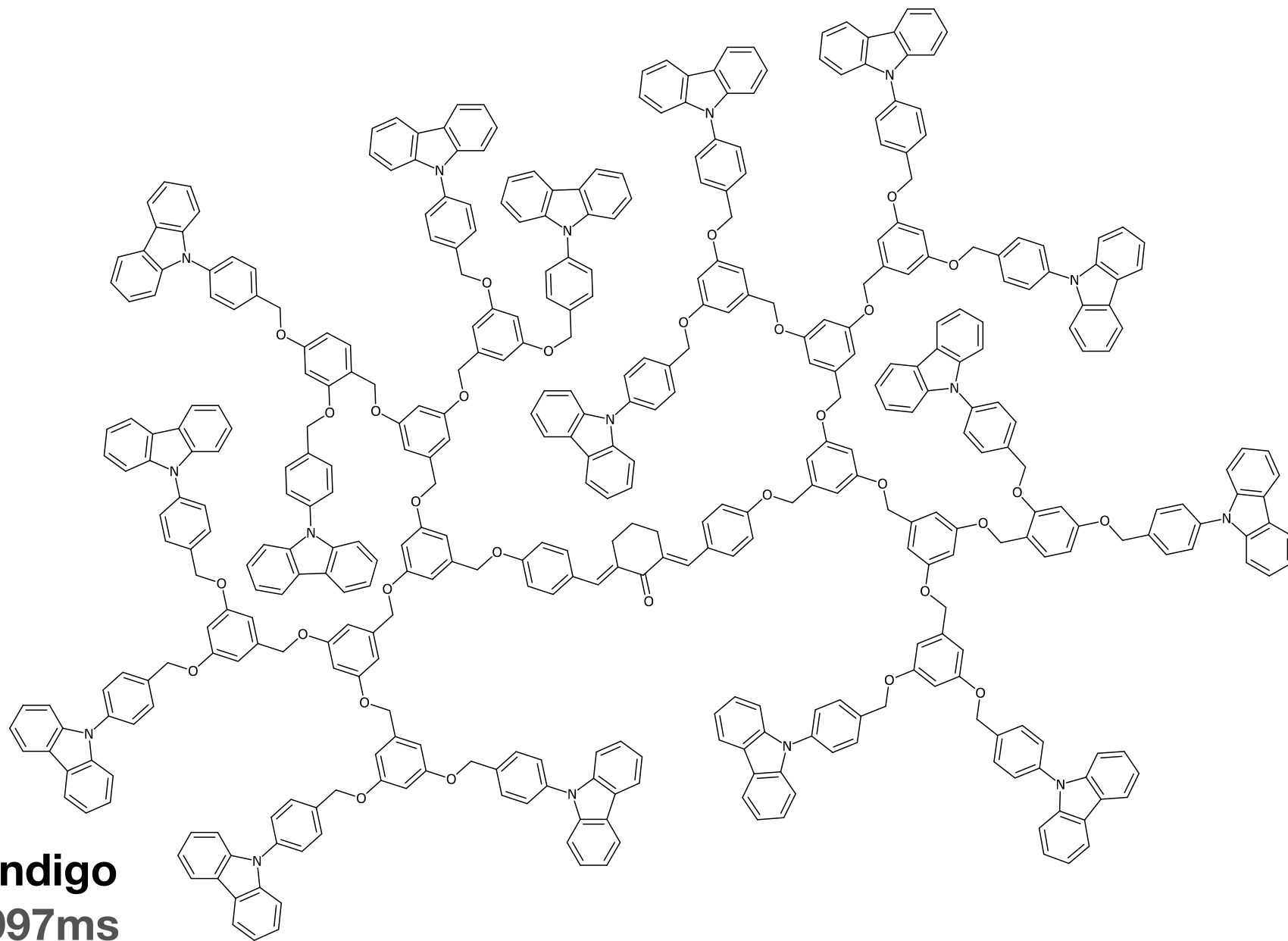
RDKit
10s

CHEMBL1077020



CDK
300ms

CHEMBL1077020



Indigo
997ms

Level 2 - Suboptimal solution

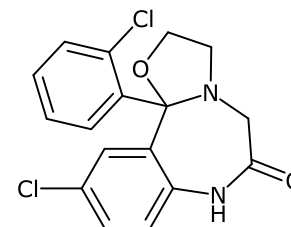
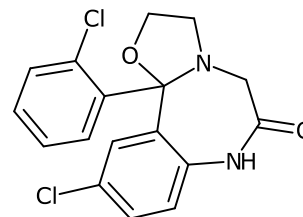
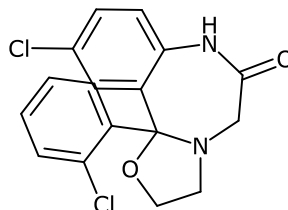
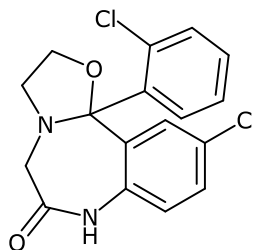
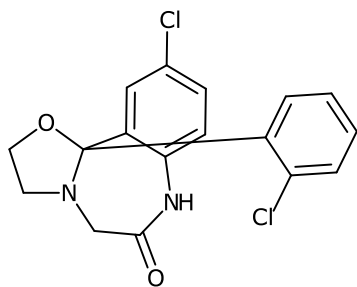
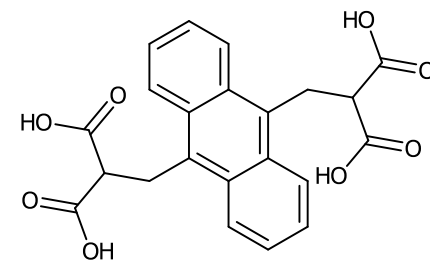
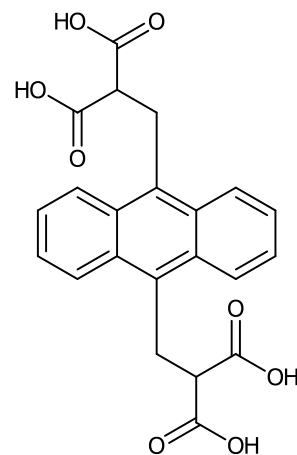
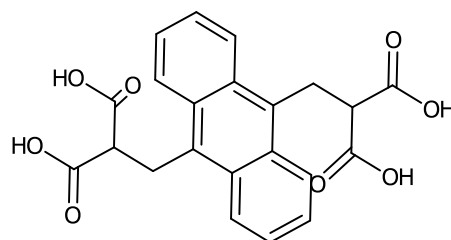
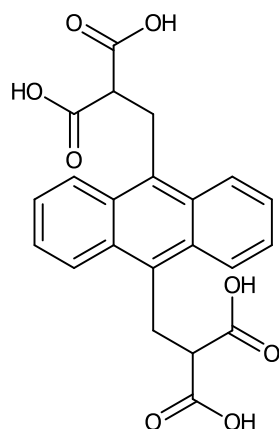
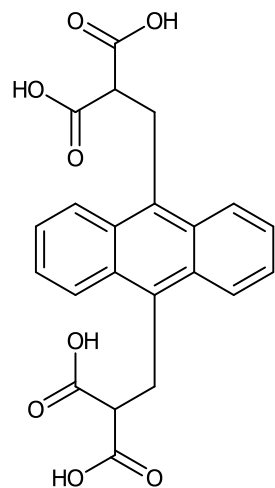
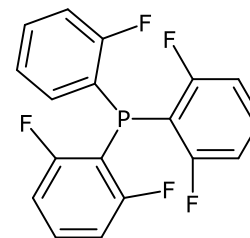
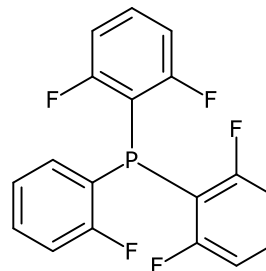
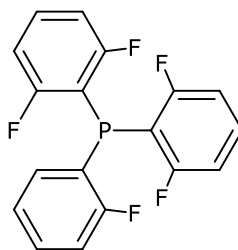
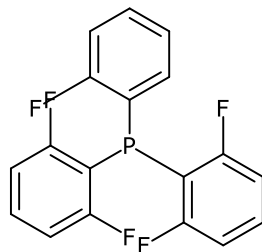
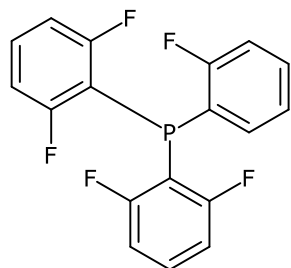
Open Babel

Avalon

RDKit

CDK

Indigo



Level 2 - Suboptimal solution

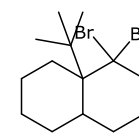
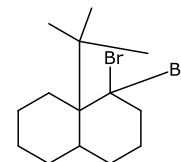
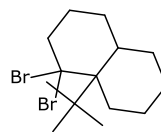
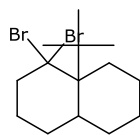
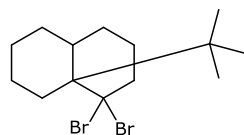
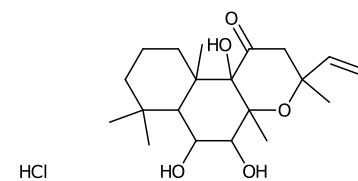
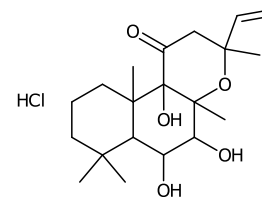
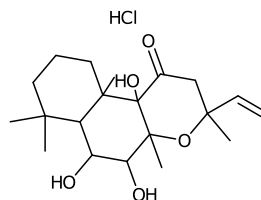
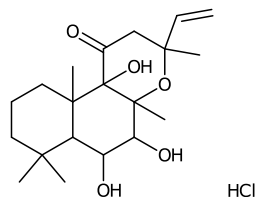
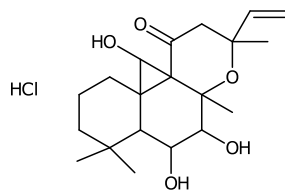
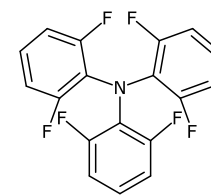
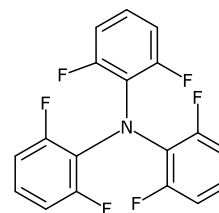
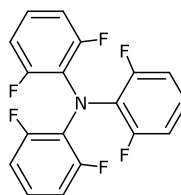
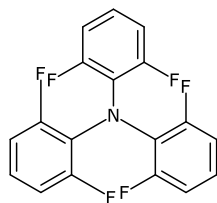
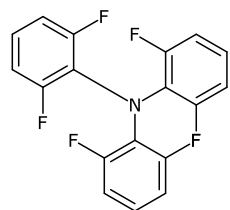
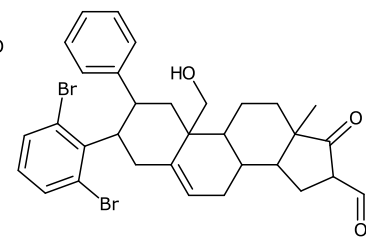
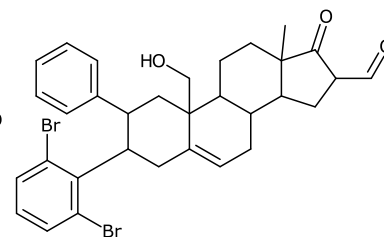
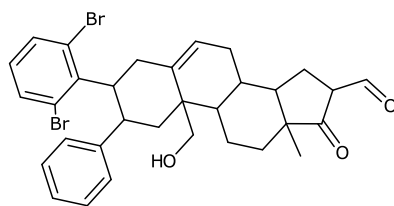
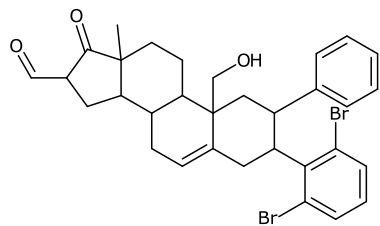
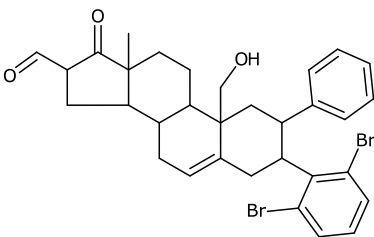
Open Babel

Avalon

RDKit

CDK

Indigo



Level 2 - Suboptimal solution

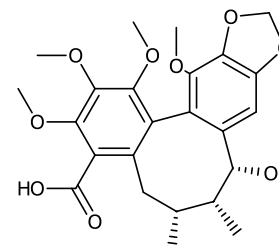
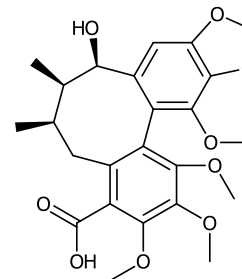
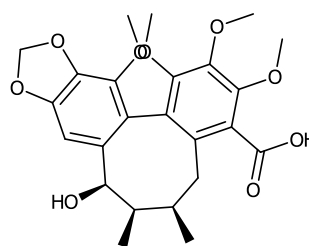
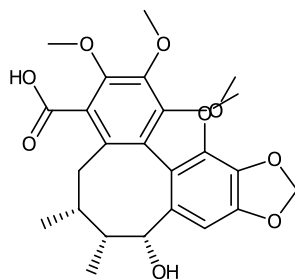
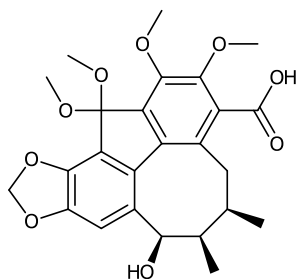
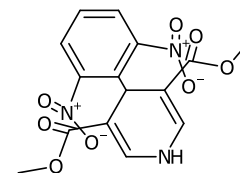
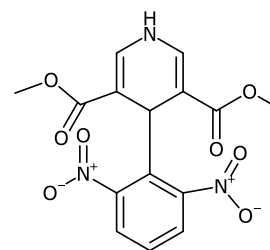
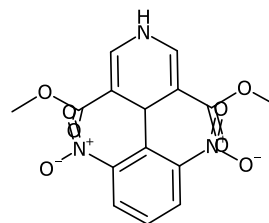
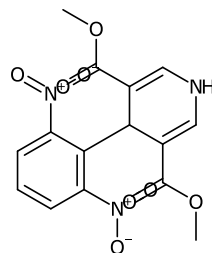
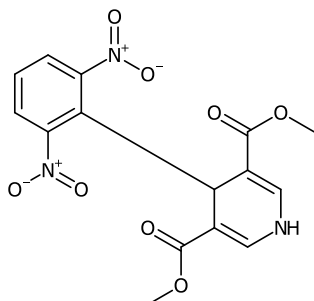
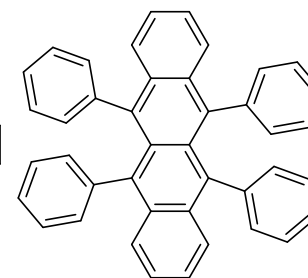
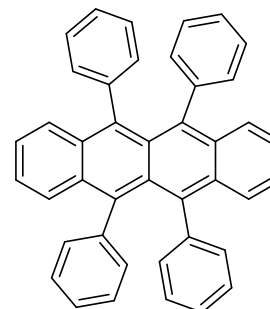
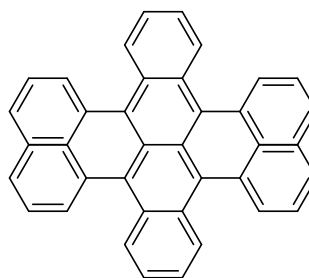
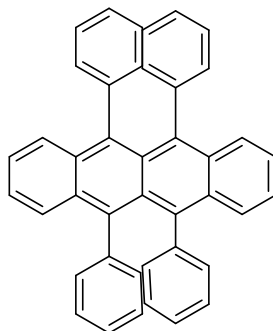
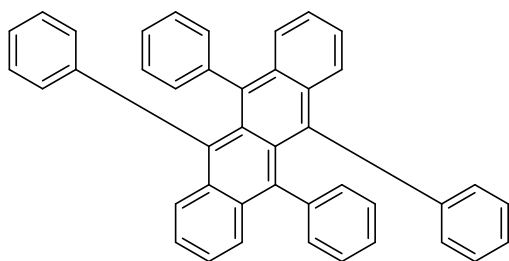
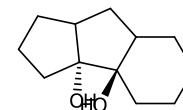
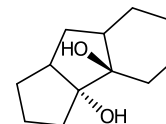
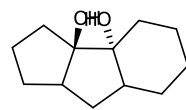
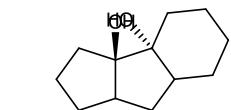
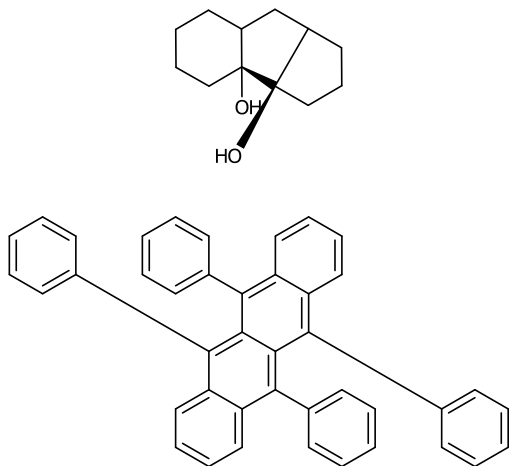
Open Babel

Avalon

RDKit

CDK

Indigo



Level 4 - Cis/Trans Bonds

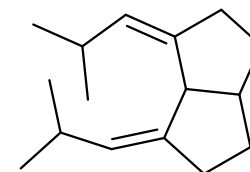
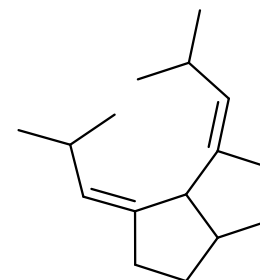
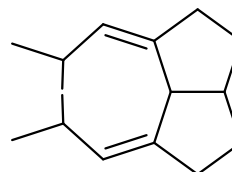
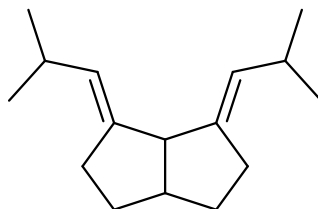
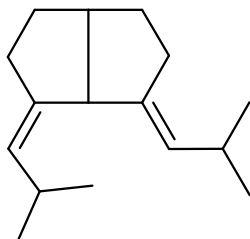
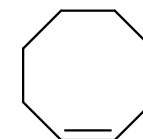
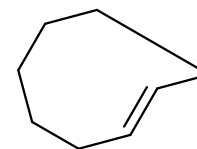
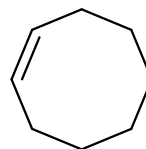
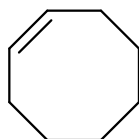
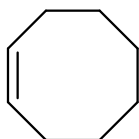
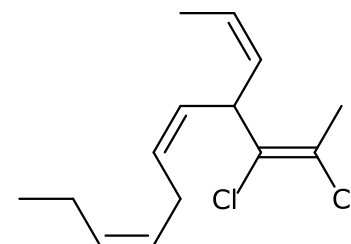
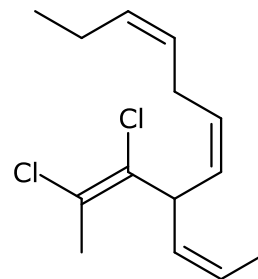
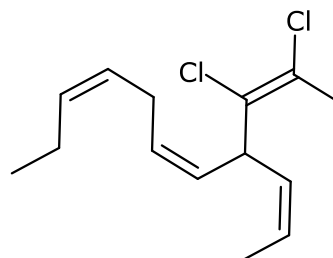
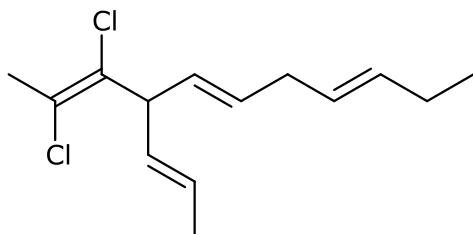
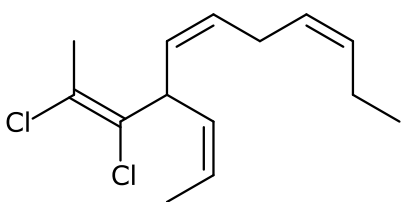
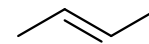
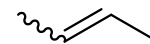
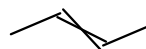
Open Babel

Avalon

RDKit

CDK

Indigo



Level 5 - Congested Rings

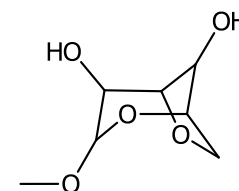
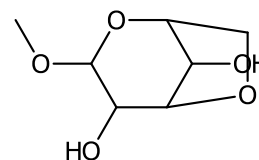
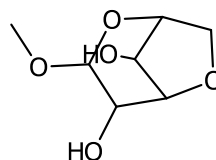
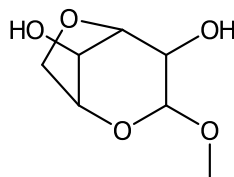
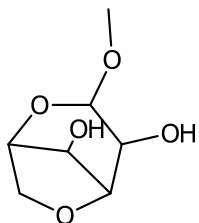
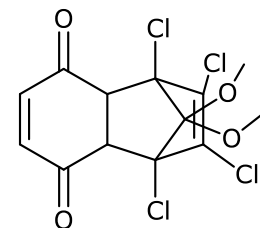
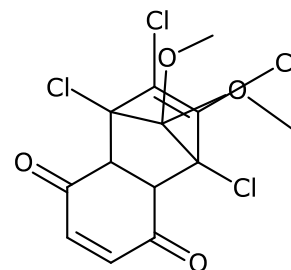
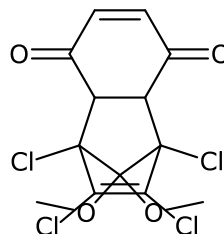
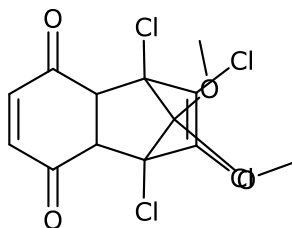
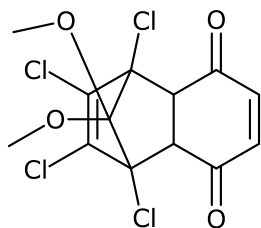
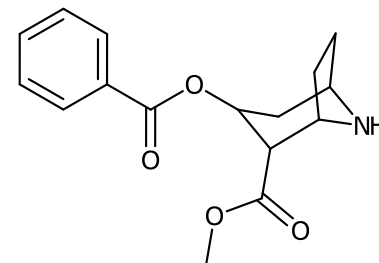
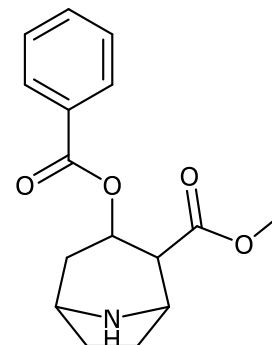
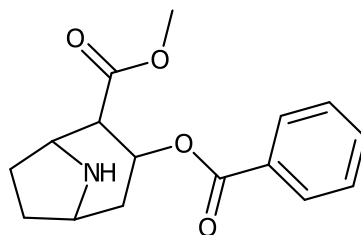
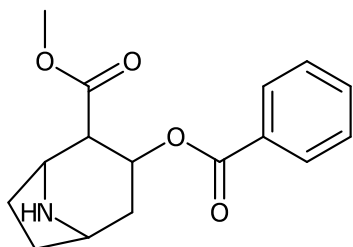
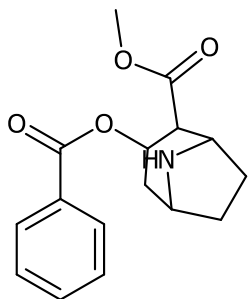
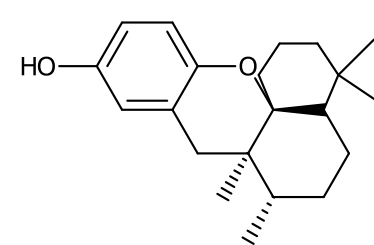
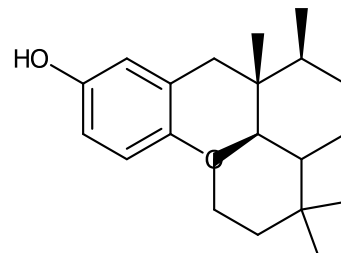
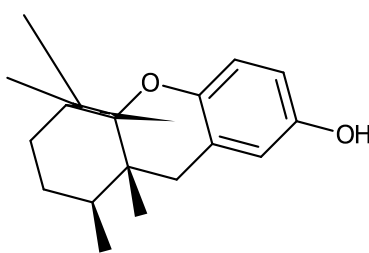
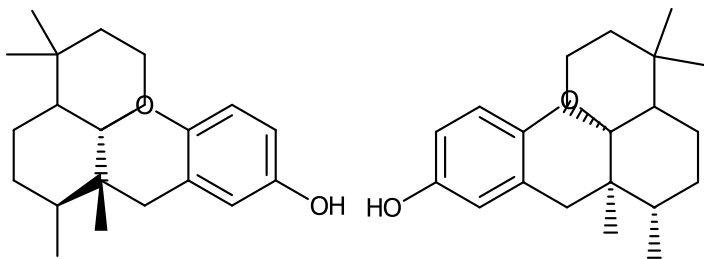
Open Babel

Avalon

RDKit

CDK

Indigo



Level 6 - Counterions

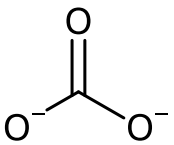
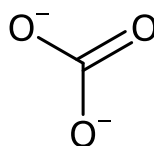
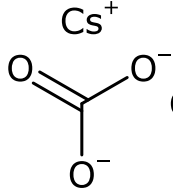
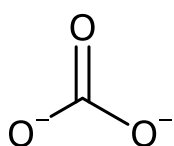
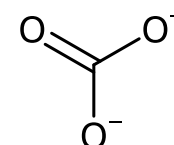
Open Babel

Avalon

RDKit

CDK

Indigo

Cs^+  Cs^+	 Cs^+ Cs^+	Cs^+  Cs^+	Cs^+  Cs^+	Cs^+  Cs^+
Al^{3+} H^+ H^- H^-	H^- H^- H^- H^- Al^{3+} Li^+	H^- H^- Li^+ Al^{3+} H^- H^-	Li^+ H^- H^- H^- Al^{3+} H^-	H^- H^- H^- Li^+ Al^{3+} H^-

Level 8 - Macrocycles

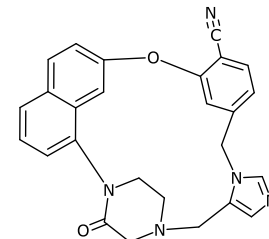
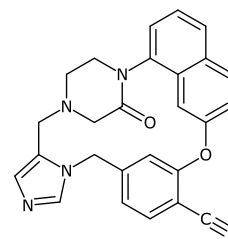
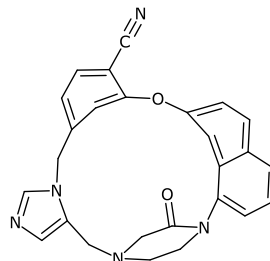
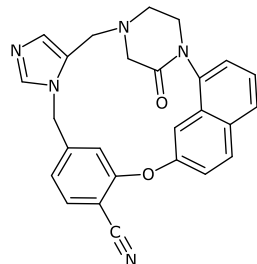
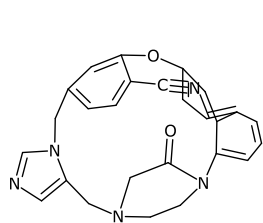
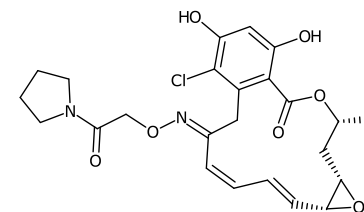
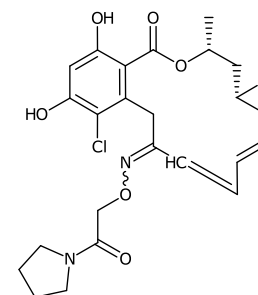
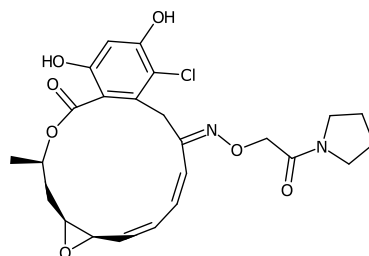
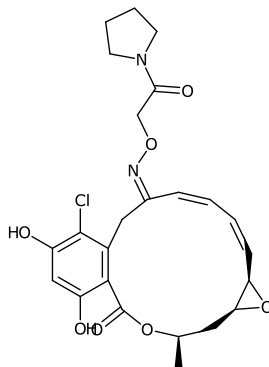
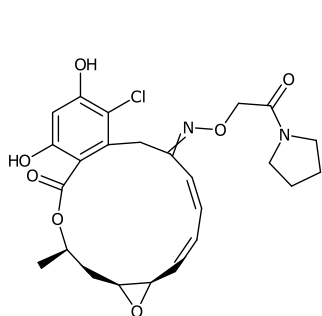
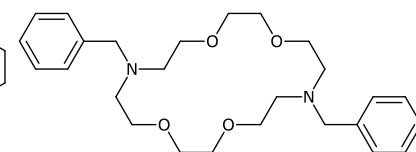
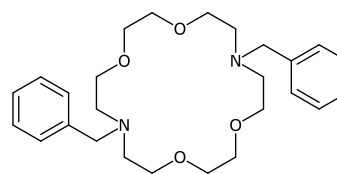
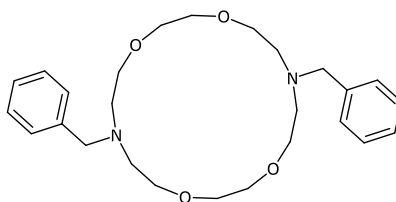
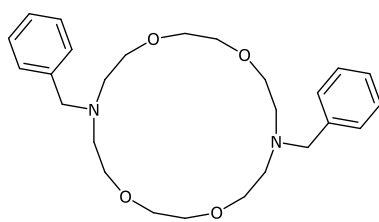
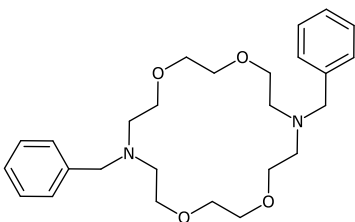
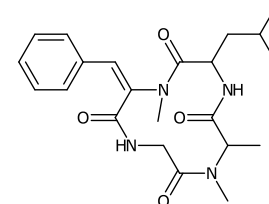
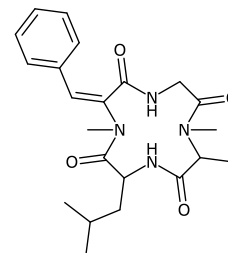
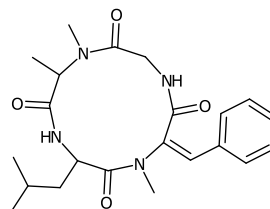
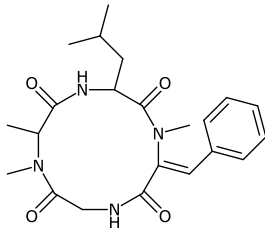
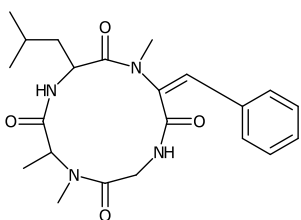
Open Babel

Avalon

RDKit

CDK

Indigo
+smart-layout



Level 8 - Macrocycles

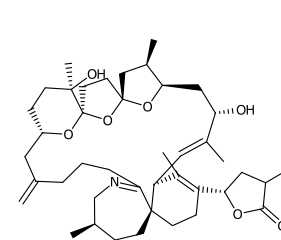
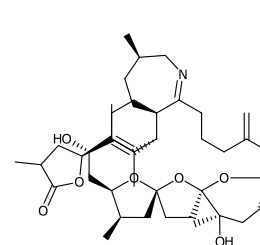
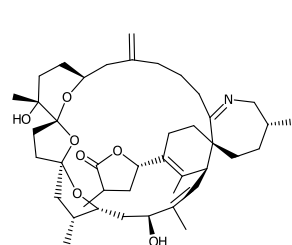
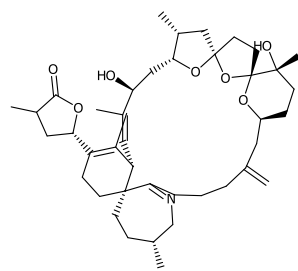
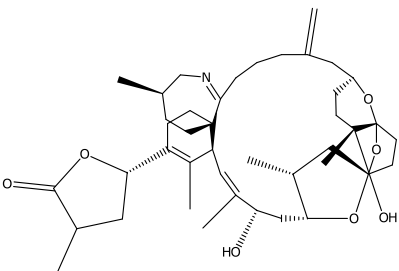
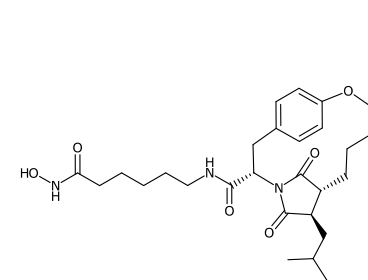
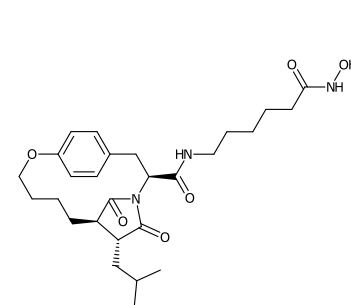
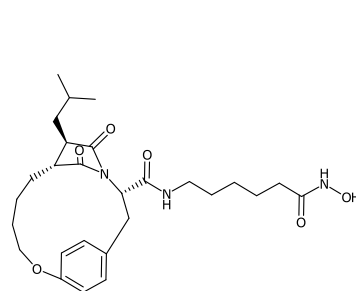
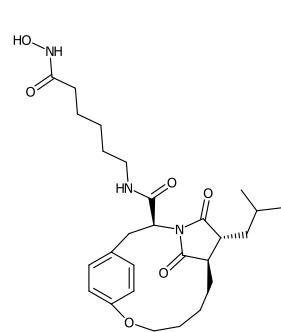
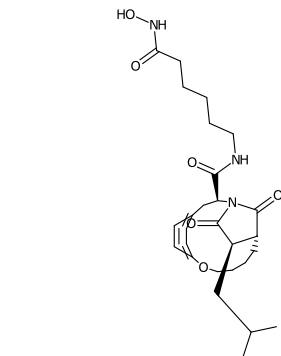
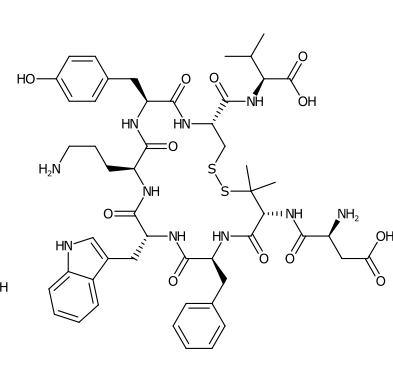
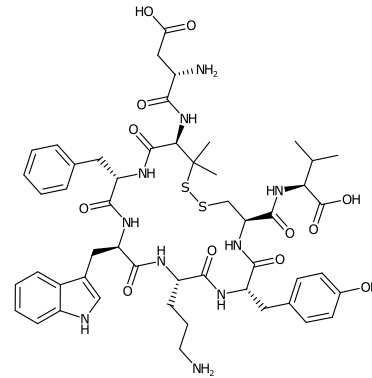
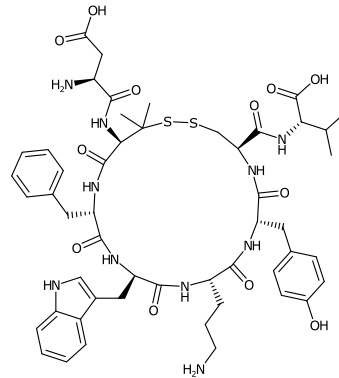
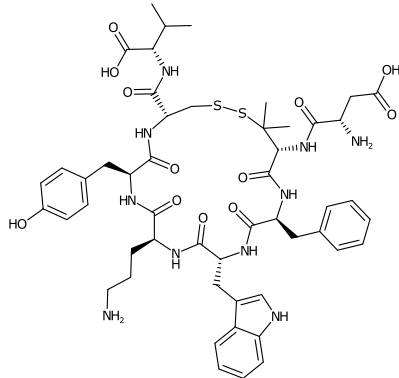
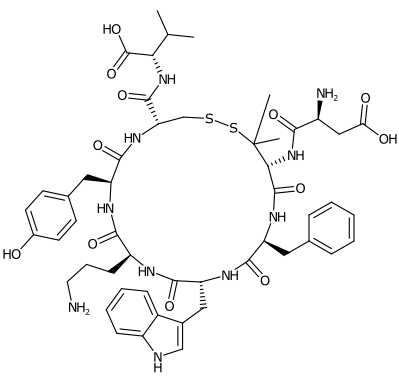
Open Babel

Avalon

RDKit

CDK

Indigo +smart-layout



Level 9/10 - Ring Template/Embedding

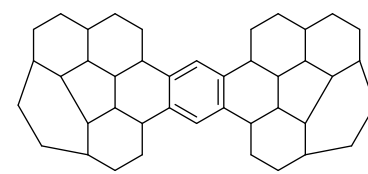
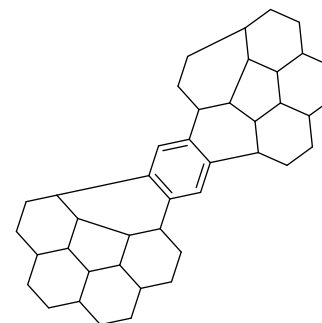
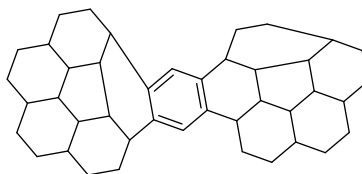
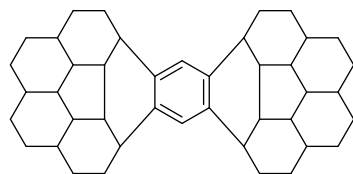
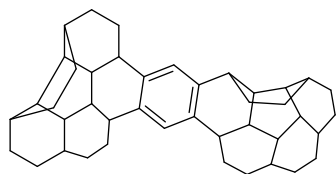
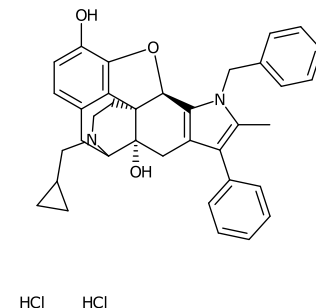
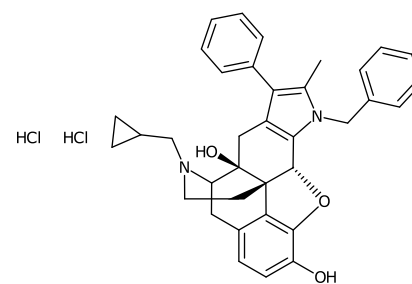
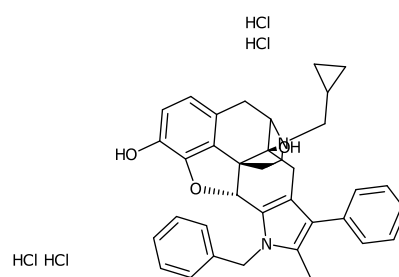
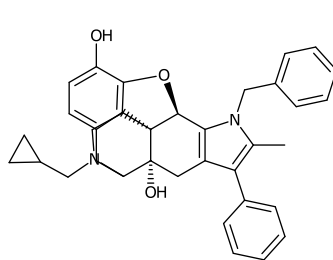
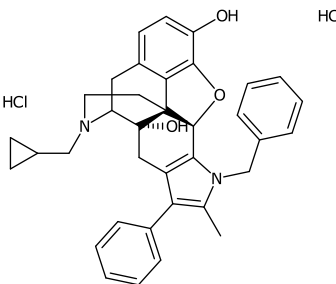
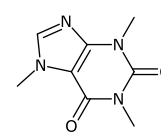
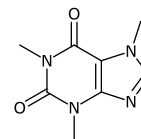
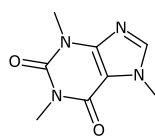
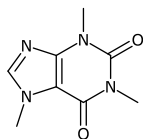
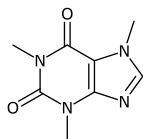
Open Babel

Avalon

RDKit

CDK

Indigo



RENDERING

A lot of quick wins for **RDKit** are in improving rendering capabilities.

Measurement and Parameters

Avoid “**Angstroms**” in layout (CDK) and drawing (RDKit), 2D depictions are not accurate or scale models!

px okay for raster but **pt**, **mm** better for vector graphics and publications

Journal Style: **ACS 1996**

Bond Spacing	18%
Bond Length	14.4pt
Bond Width	2pt
Line Width	0.6pt
Margin Width	1.6pt
Hash Spacing	2.5pt
Captions	10pt
Atom Labels	10pt

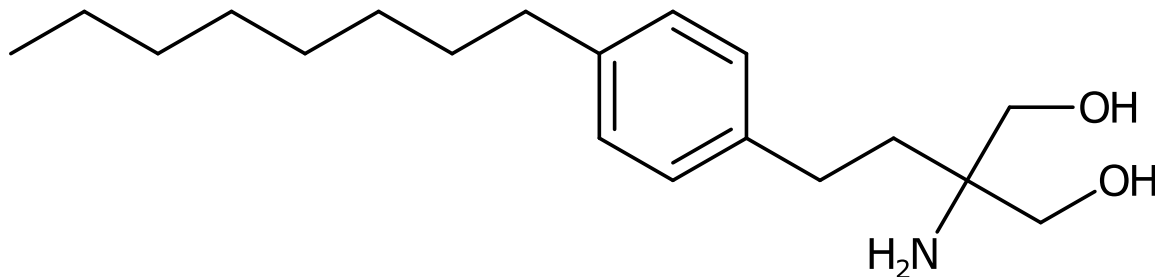
... π bond width

...wedge bond width

...annotations



How Many `<text>` Elements in SVG?



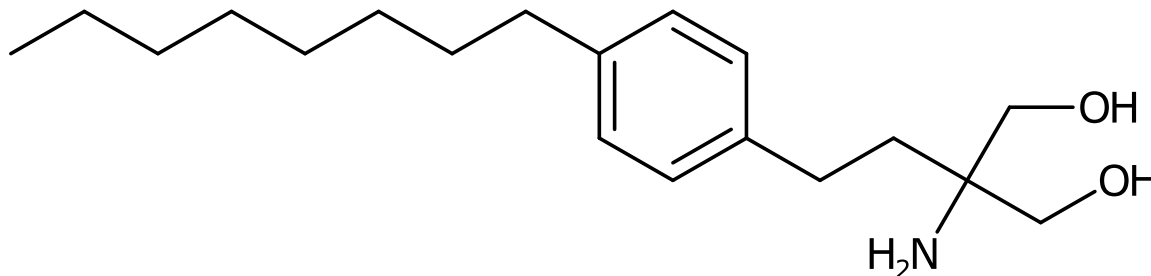
(a) 3

(b) 6

(c) 7



How Many <text> Elements?

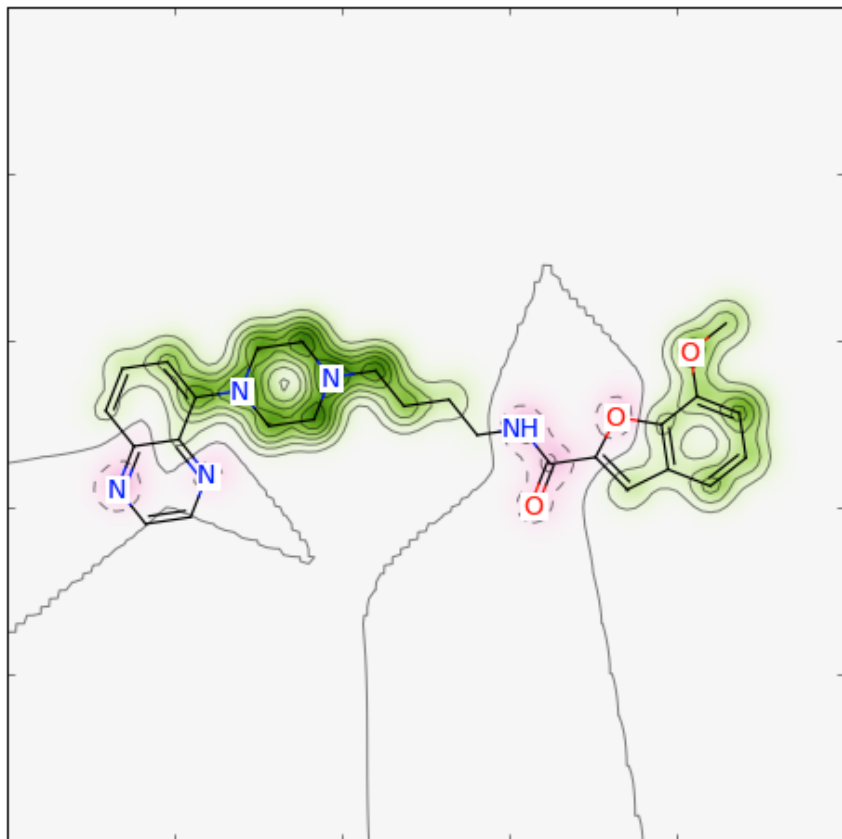


- ~~(a) 3~~
- ~~(b) 6~~
- ~~(c) 7~~
- (d) 0**

```
<g id='mol1atm19' class='atom'>
  <path d='M67.32 10.16q.11 .12 .17 .3q.06 .18 .06 .41q.0 .23 -.06 .41q-.06 .18 -.17 .3q-.1
  -.36 .06q-.19 .0 -.35 -.06q-.16 -.06 -.27 -.19q-.11 -.12 -.17 -.3q-.06 -.18 -.06 -.41q.0
  -.31q.11 -.12 .27 -.19q.16 -.06 .35 -.06q.2 .0 .36 .07q.16 .07 .27 .19zM67.3 10.87q.0 -.3
  -.2q-.28 .0 -.44 .2q-.16 .2 -.16 .56q.0 .36 .16 .56q.16 .19 .44 .19q.28 .0 .44 -.19q.16 -
  />
  <path d='M69.18 11.79h-.25v-.9h-.92v.9h-.25v-1.85h.25v.72h.92v-.72h.25z' stroke='none' />
</g>
<g id='mol1atm21' class='atom'>
  <path d='M56.06 16.19h-.3l-.88 -1.65v1.65h-.23v-1.85h.38l.8 1.51v-1.51h.23z' stroke='none'
  <path d='M53.48 16.19h-.25v-.9h-.92v.9h-.25v-1.85h.25v.72h.92v-.72h.25z' stroke='none' />
  <path d='M54.44 16.75h-.75v-.16q.08 -.07 .16 -.13q.08 -.07 .15 -.13q.14 -.14 .2 -.22q.05
  -.14q-.06 -.05 -.16 -.05q-.07 .0 -.15 .02q-.08 .02 -.15 .07v.0v-.16q.05 -.03 .14 -.05q.05
  .08q.1 .08 .1 .22q.0 .06 -.02 .12q-.02 .05 -.05 .1q-.03 .05 -.07 .09q-.04 .04 -.09 .1q-.6
  -.16 .14h.6z' stroke='none' />
</g>
<g id='mol1atm22' class='atom'>
  <path d='M66.14 5.76q.11 .12 .17 .3q.06 .18 .06 .41q.0 .23 -.06 .41q-.06 .18 -.17 .3q-.1
  .06q-.19 .0 -.35 -.06q-.16 -.06 -.27 -.19q-.11 -.12 -.17 -.3q-.06 -.18 -.06 -.41q.0 -.23
  -.31q.11 -.12 .27 -.19q.16 -.06 .35 -.06q.2 .0 .36 .07q.16 .07 .27 .19zM66.12 6.47q.0 -.3
  -.2q-.28 .0 -.44 .2q-.16 .2 -.16 .56q.0 .36 .16 .56q.16 .19 .44 .19q.28 .0 .44 -.19q.16 -
  />
  <path d='M68.0 7.39h-.25v-.9h-.92v.9h-.25v-1.85h.25v.72h.92v-.72h.25z' stroke='none' />
</g>
```



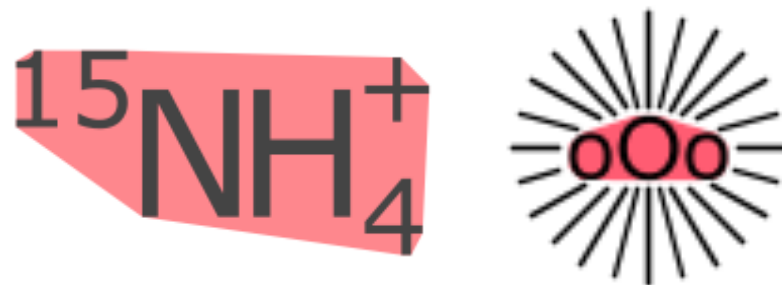
Font Embedding



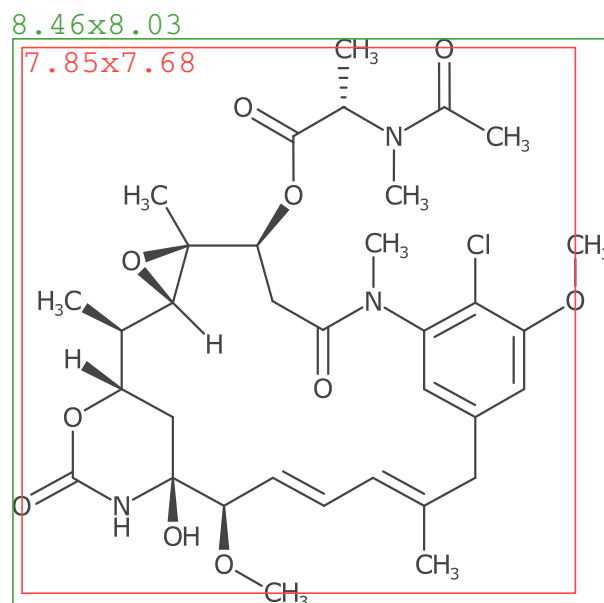
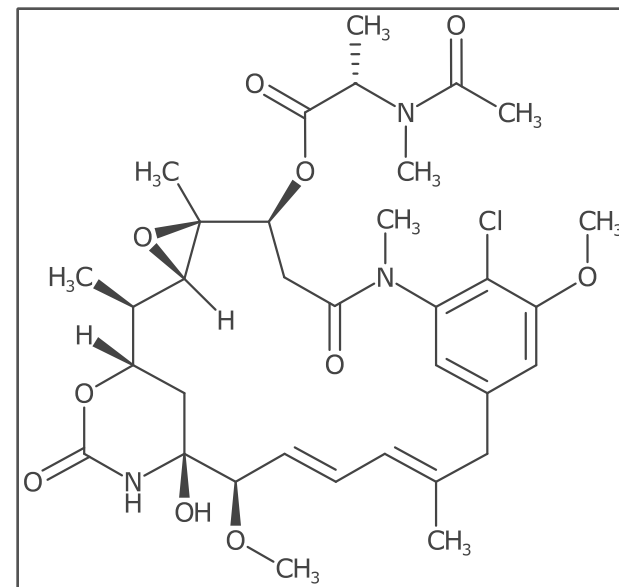
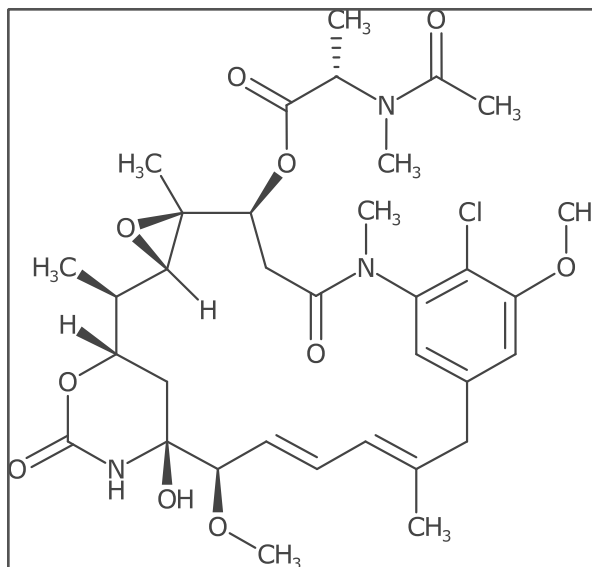
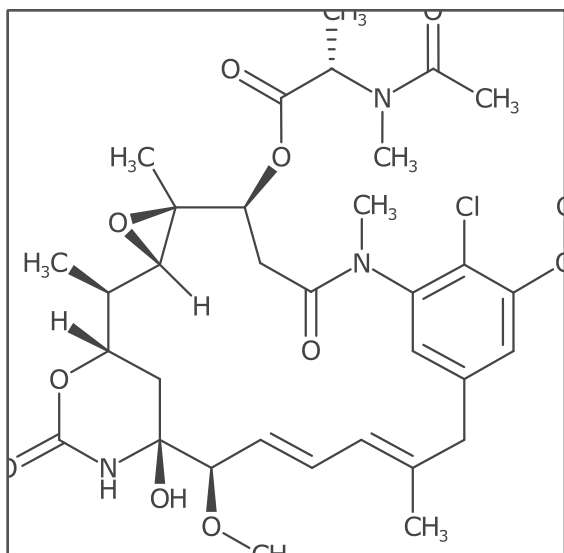
More Portable



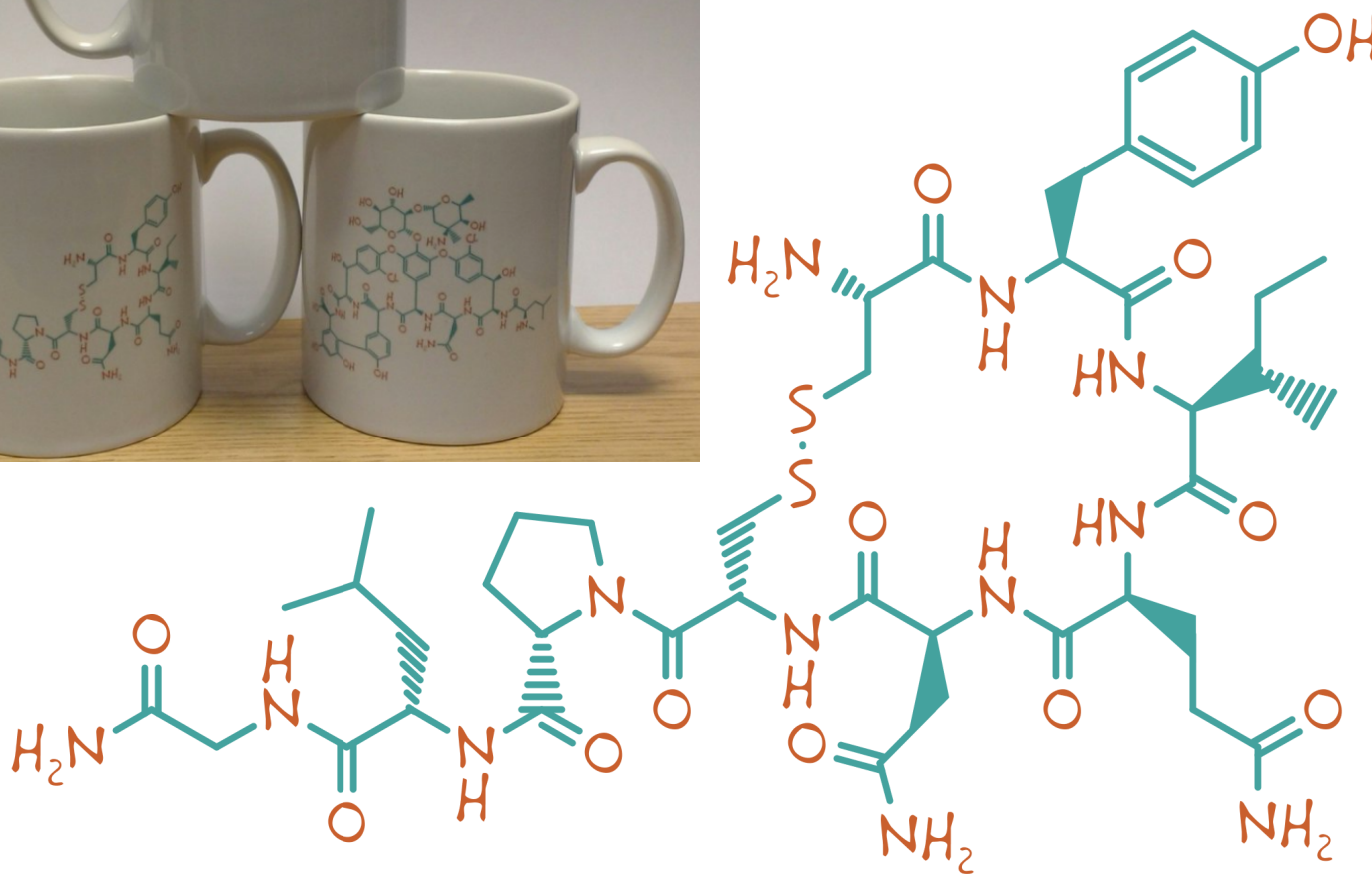
Convex Hull Bounds



Bounding Box

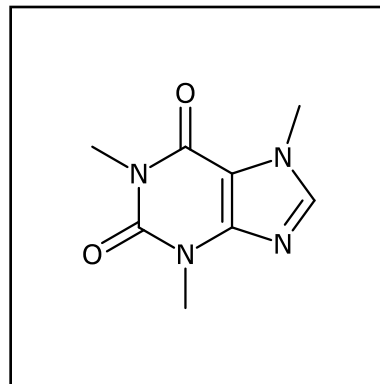
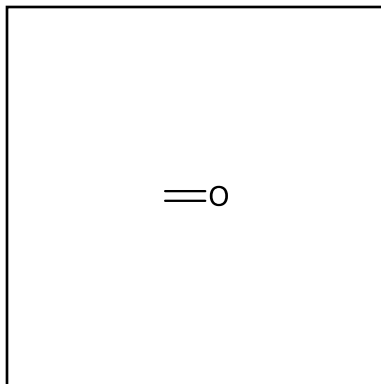
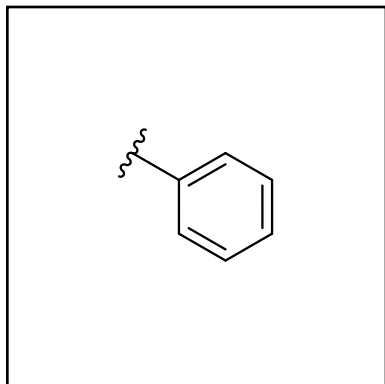
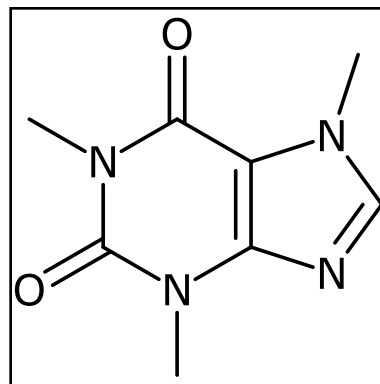
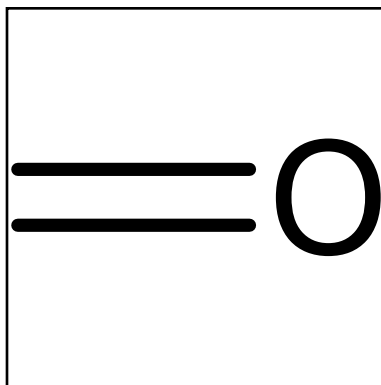
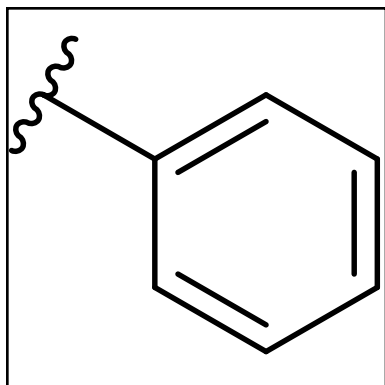


FUN WITH FONTS



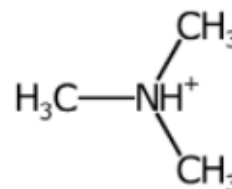
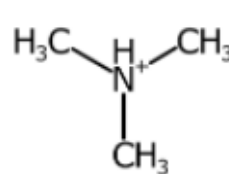
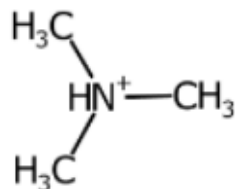
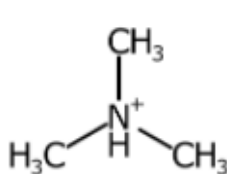
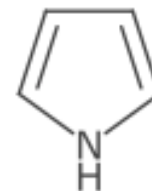
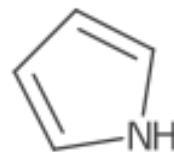
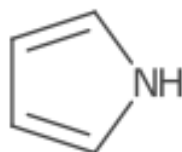
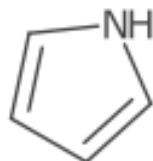
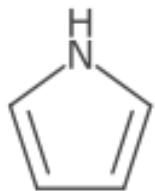
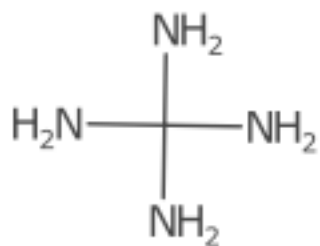
Depiction Scale

Control depiction size by **bond length** parameter. Shrink to fit, but avoid **stretch to fit** (make optional?).

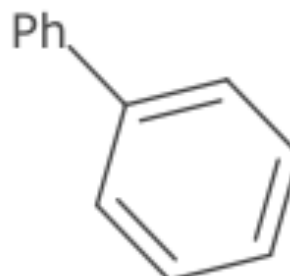
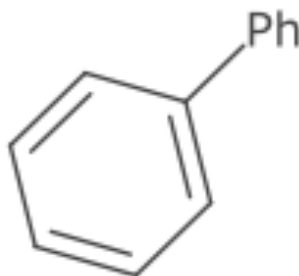
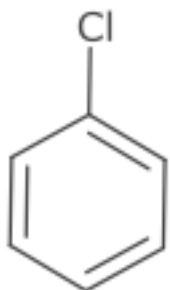


Adjunct Placement and Alignment

Hydrogen Placement

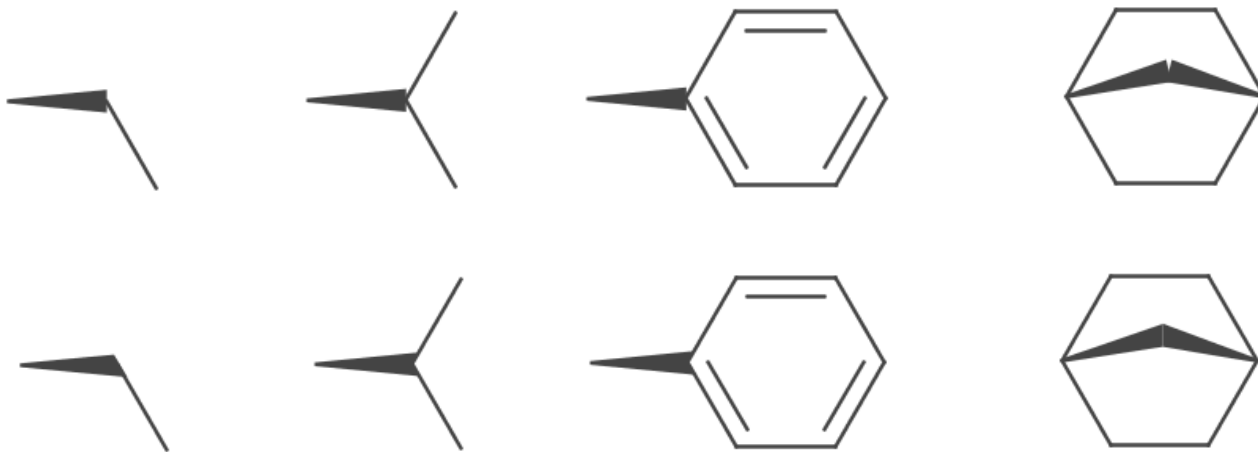


Alignment



Bold and Hashed Wedges

Slanting and Bifurcation of Wedges

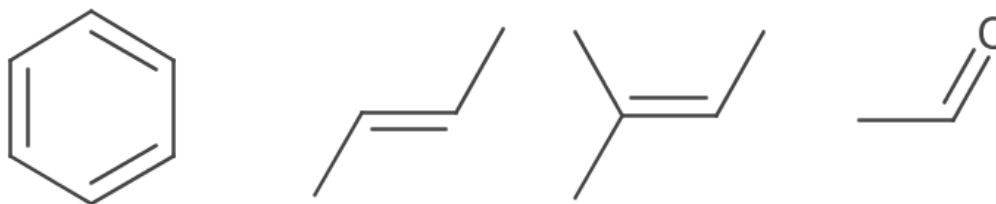


Possible with hashes but controversial

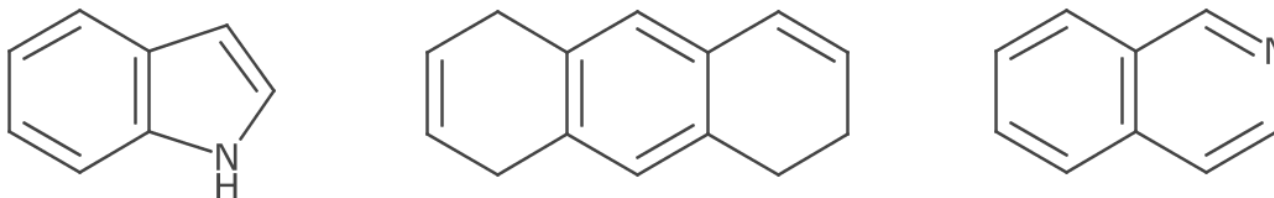


Double Bonds

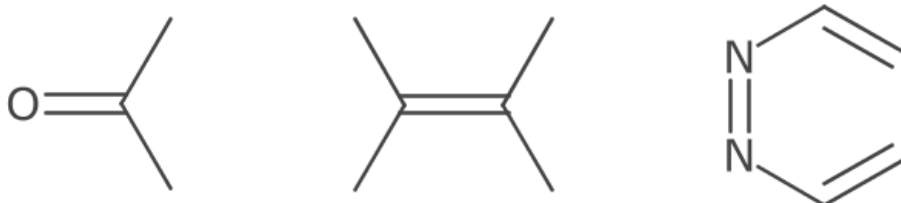
Offset Double Bonds



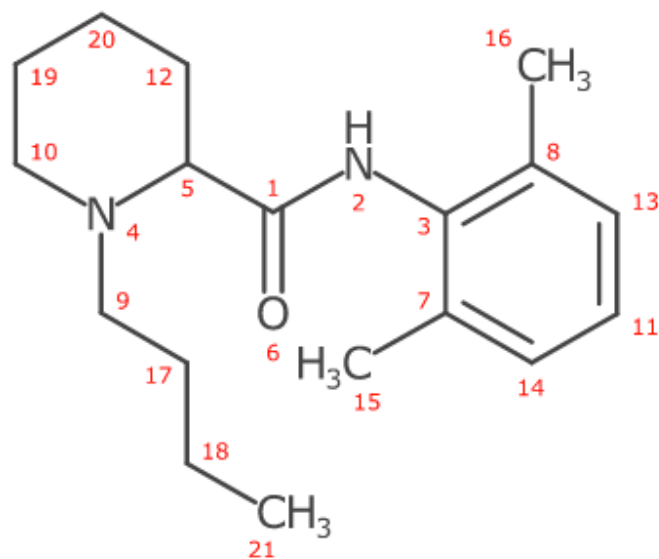
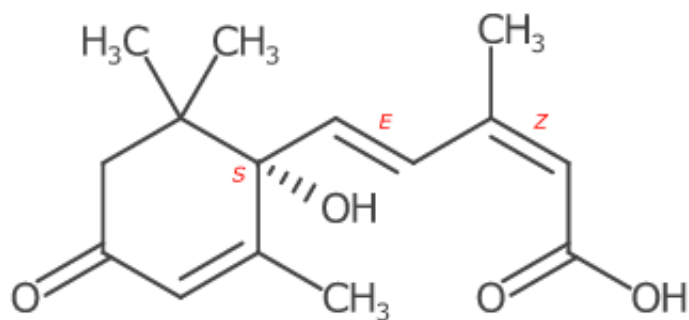
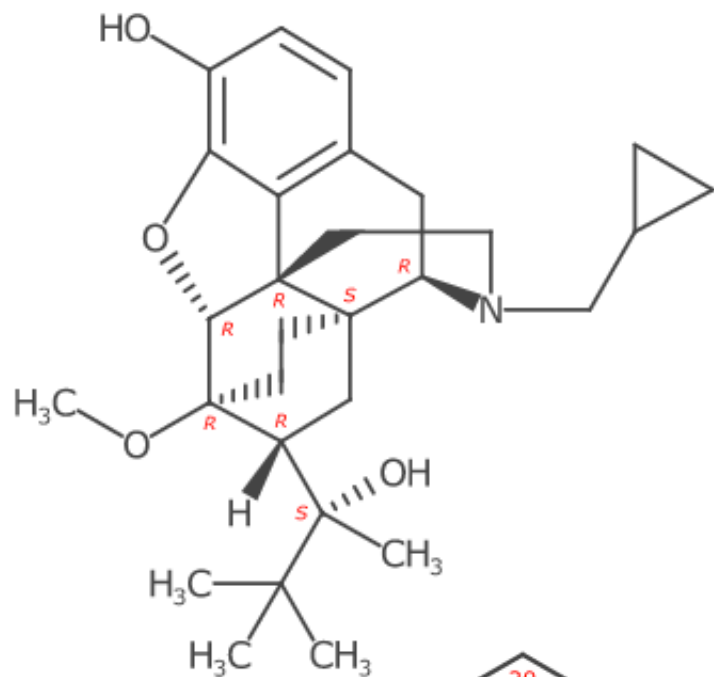
Which side? (General rule=more benzene!)



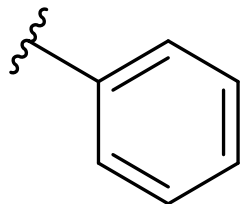
Centred Double Bonds



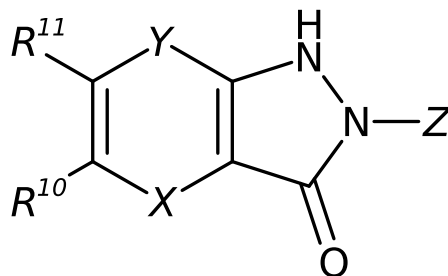
Annotations



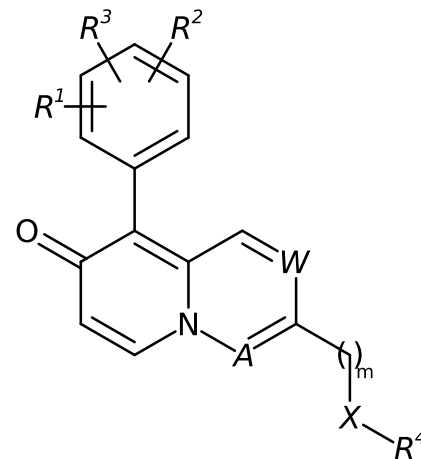
Sgroups and Generic Features



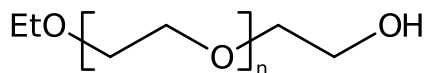
Attachment Points



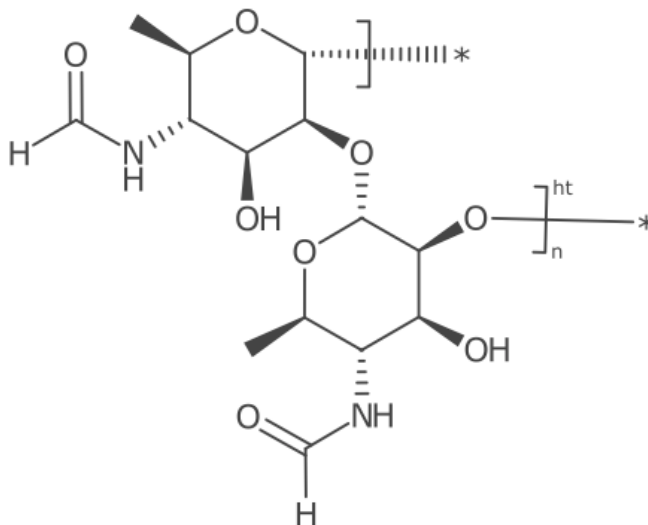
Substituent Labels
"R" Groups



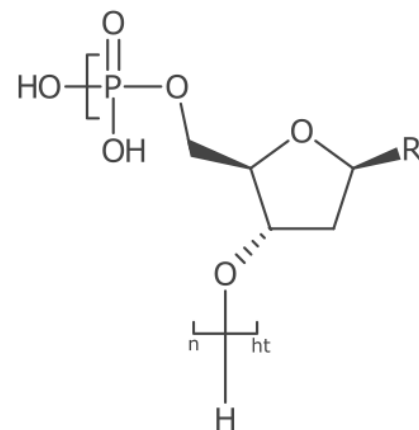
Positional Variation



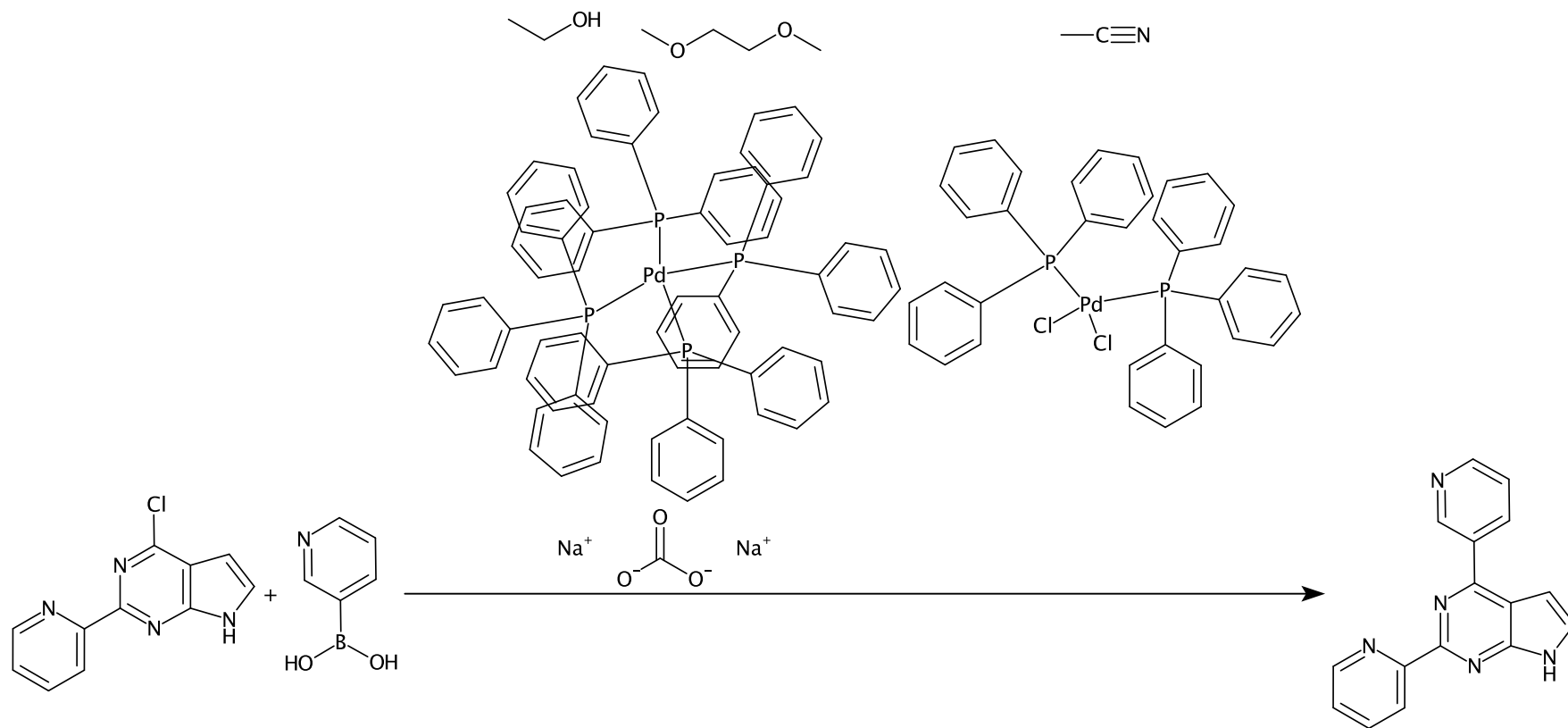
Abbreviations



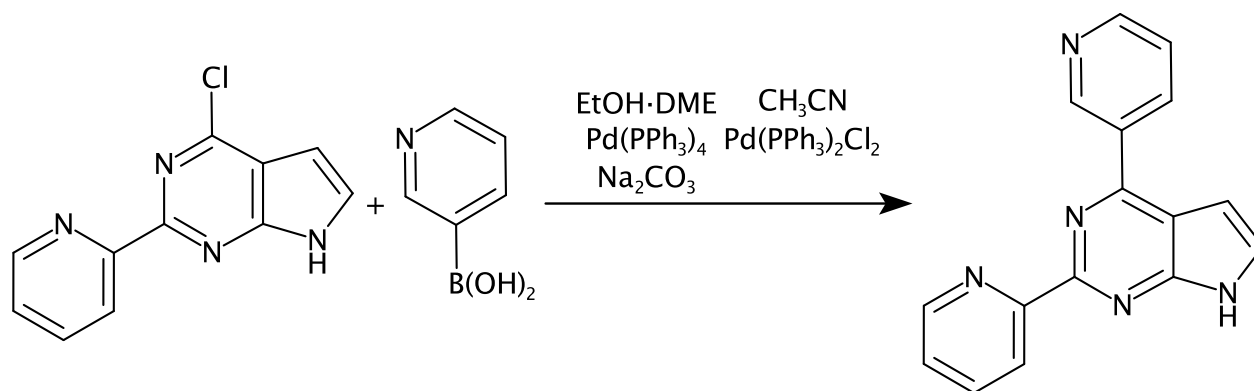
Structure Repeat Units



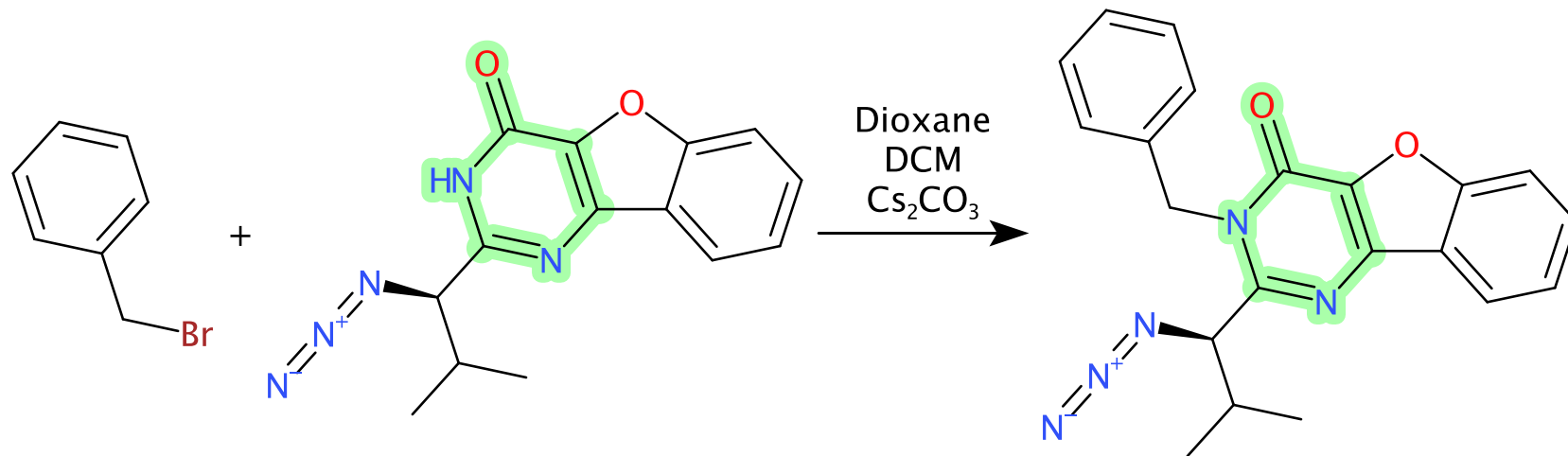
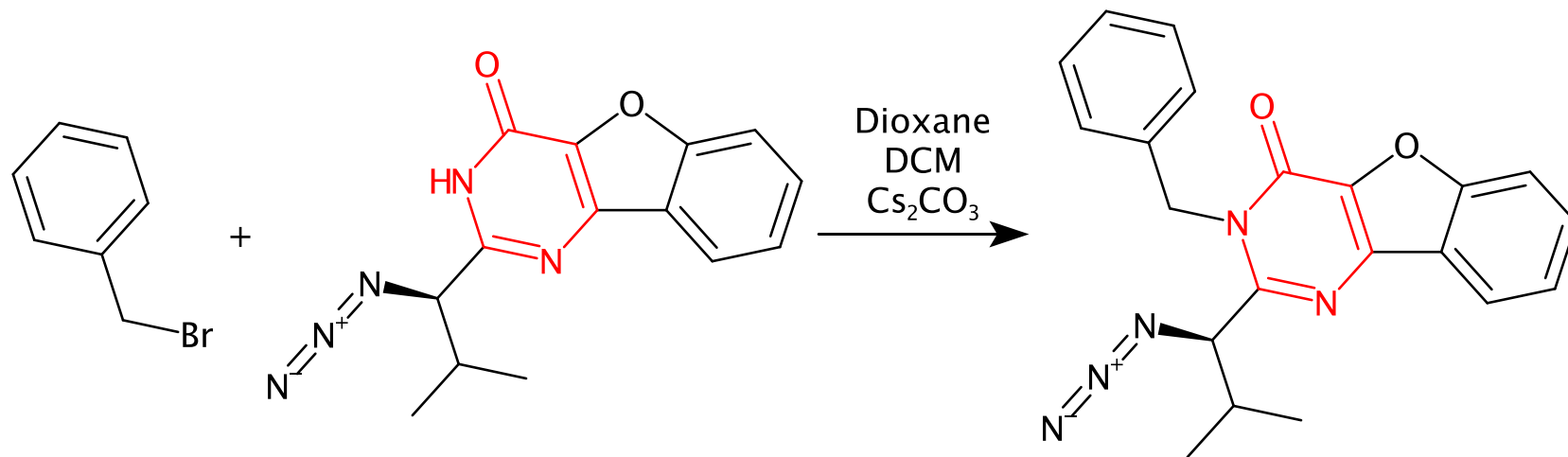
Abbreviations in Action

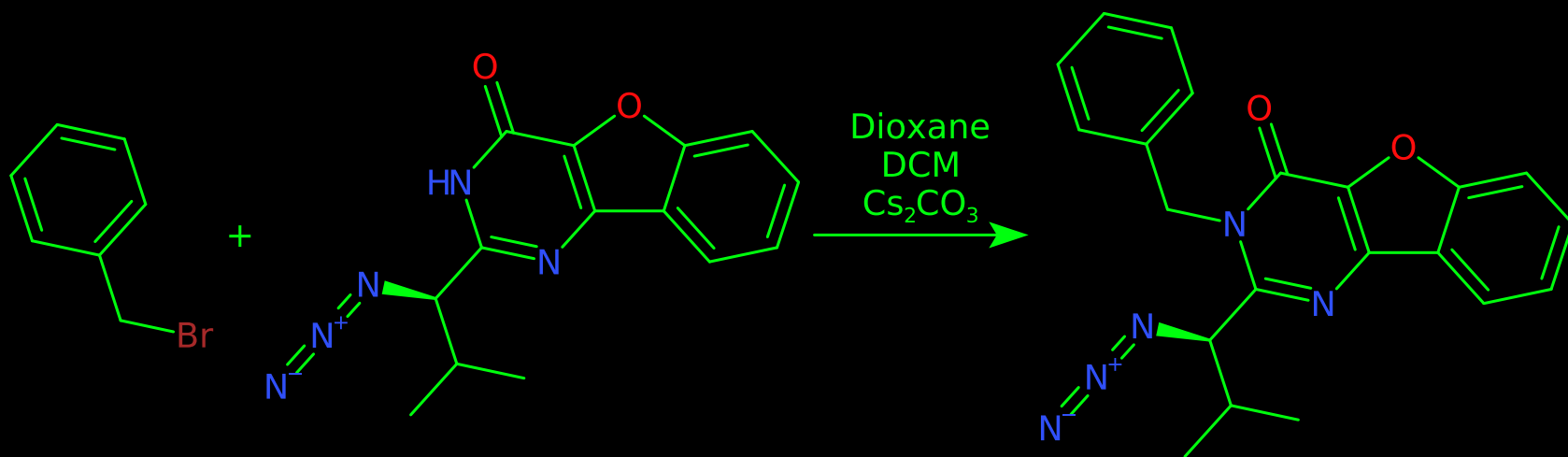


Abbreviations in Action



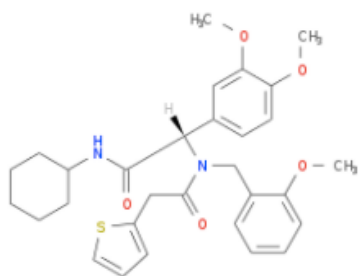
Colors



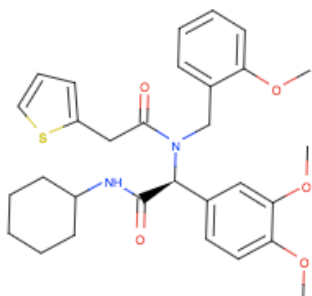


Layout + Rendering Comparison

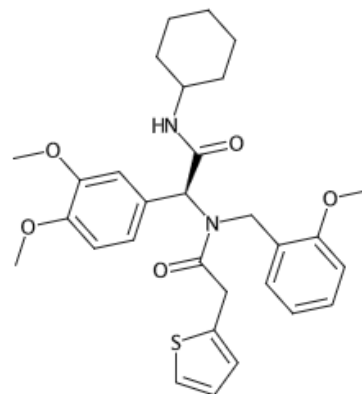
Open Babel



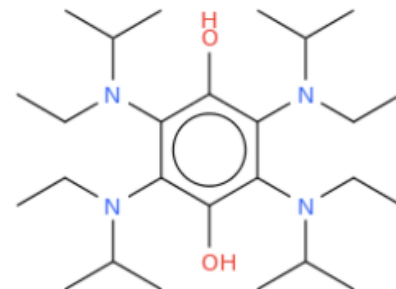
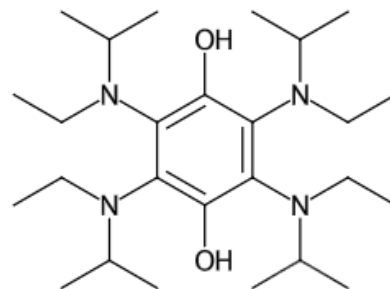
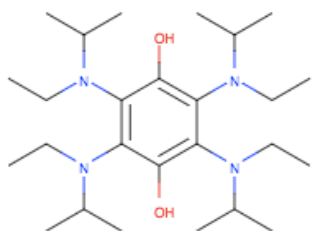
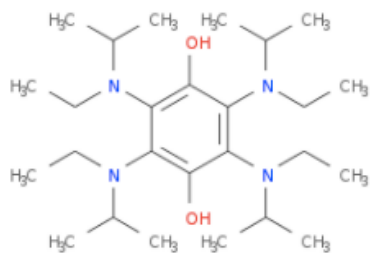
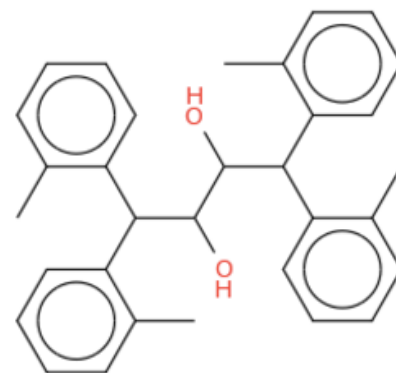
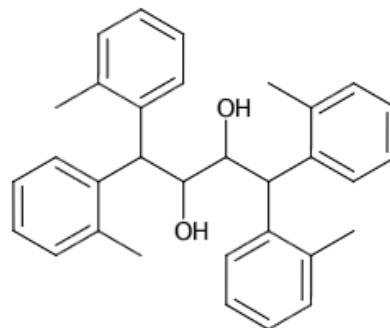
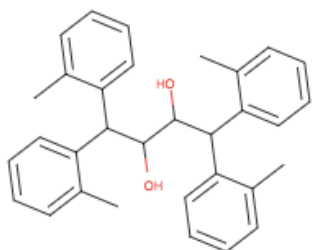
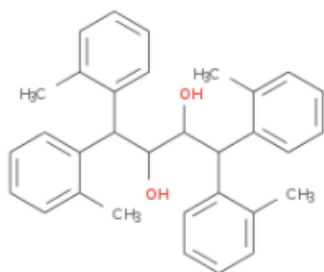
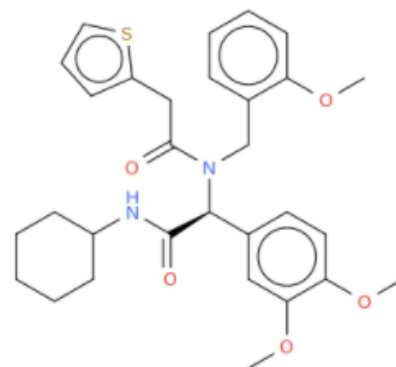
RDKit



CDK

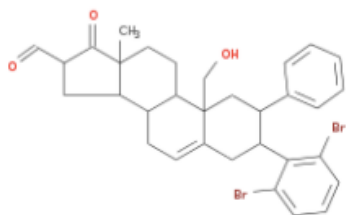


Indigo

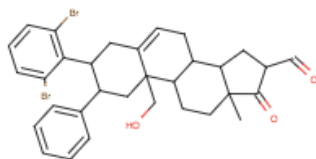


Layout + Rendering Comparison

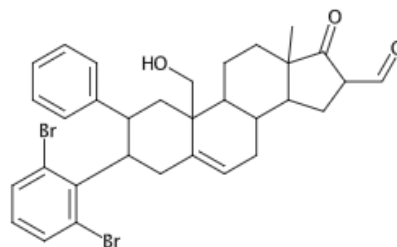
Open Babel



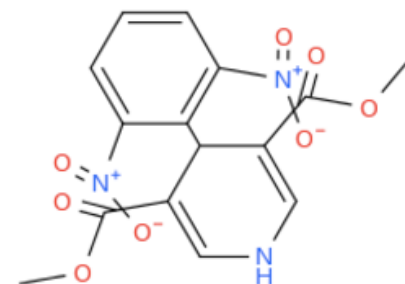
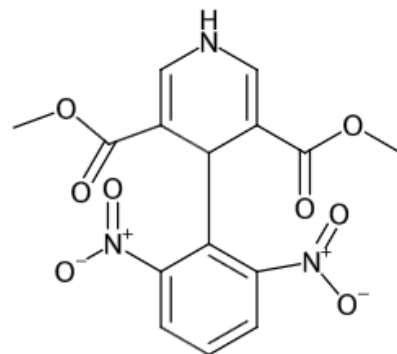
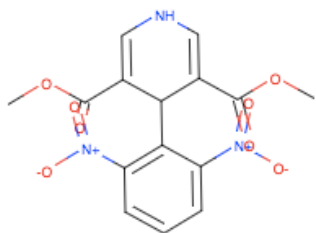
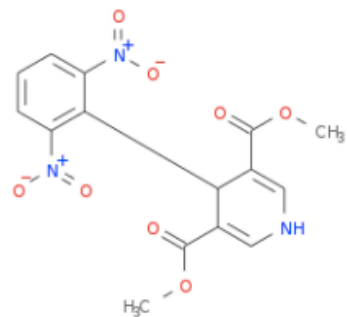
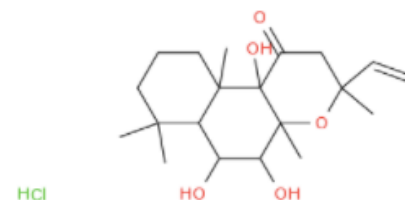
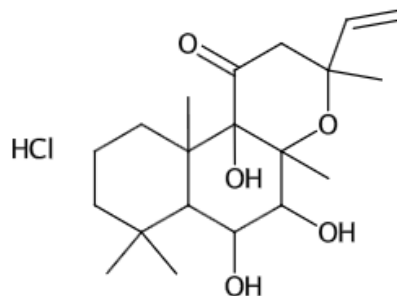
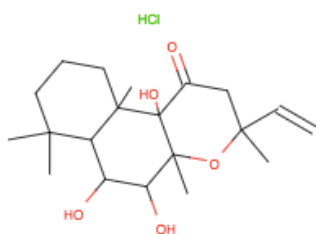
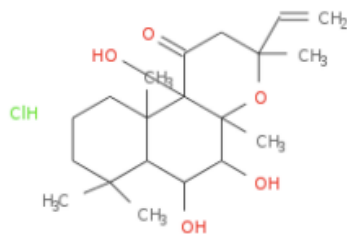
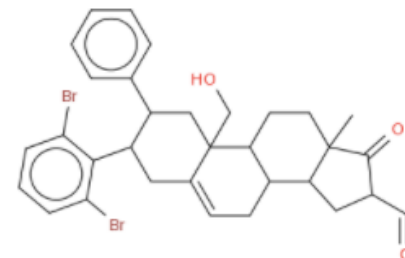
RDKit



CDK



Indigo



Acknowledgements

Constructive criticism of depictions: **Roger Sayle, Daniel Lowe, Noel O'Boyle**

Reviewing patches: **Egon Willighagen**

Initial CDK layout: **Christoph Steinbeck**

Seminal Papers: **Alex Clark** and **Jonathan Brecher**.



Spot the difference

RDKit

CDK

