Australian Government
IP Australia

AI tools in ST.96 implementation

Presented by

Markus Klaiber

International ICT Cooperation Specialist

&

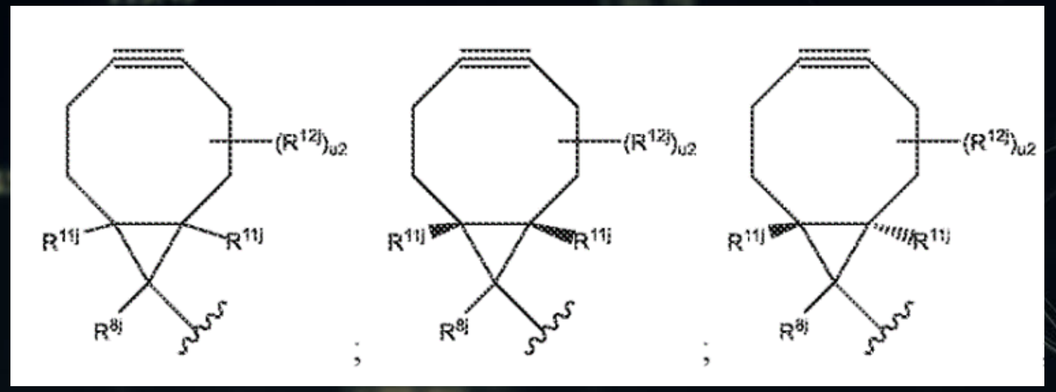
Alexander Ivakov

Lead Developer, Structured Patent Specifications

P TM D PBR

ST.96 at IP Australia

- All patent filing and processing using PDF
- Initial priority was on text quality
- OCR has come a long way, state of the art, highly accurate on blocks of text
- Problematic output from OCR mostly came from within chem, math and figures, not text



OCR

```
(RI2 J)ii2  (R19 u2 (R~2J)
a, It
* a
26
a
```

Chemical structure detection

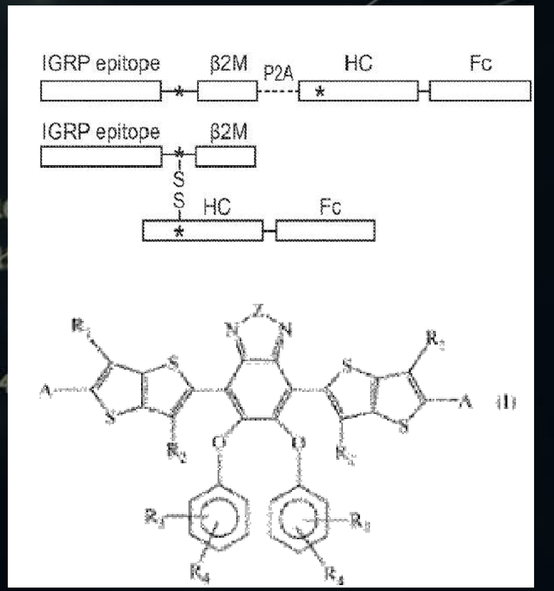
PDF

Textract

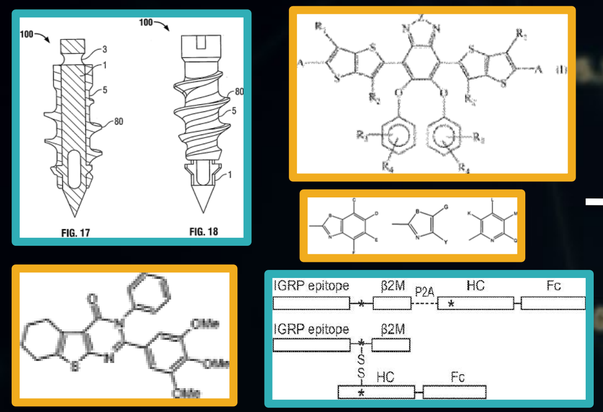
Text

[0003] Electronic devices such as laptops, tablets, or mobile phones are often used in a working environment. Typically, electronic devices are placed on support structures. It is often desired by users to adjust the height or location of the electronic device relative to the desktop to gain a better view of the electronic device. However, many support structures for electronic devices are difficult to adjust and require great effort by the user to adjust the height or location of the electronic device. Further, many support structures do not allow for a range of movement required for a full ergonomic solution and take up a large amount of space on the desktop.

Figures



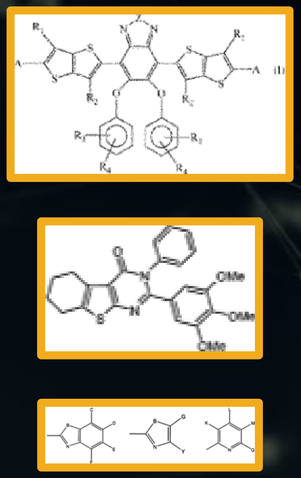
Chemical structure detection (cont'd)



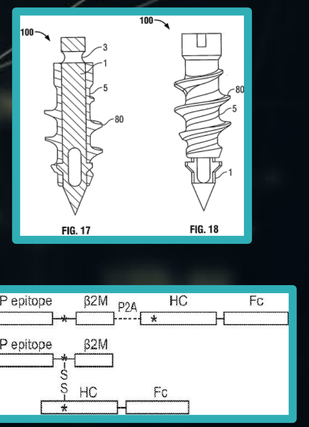
Cropped figures from Textract

Image Classifier

Chemical



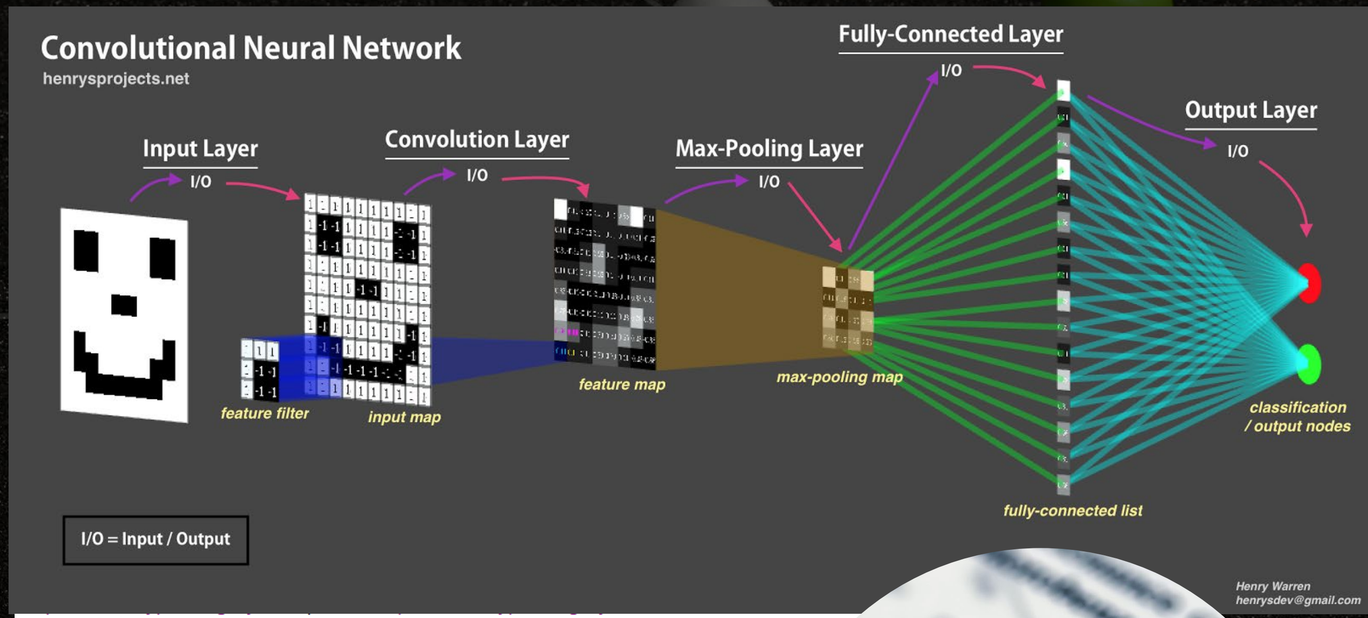
Non-chemical

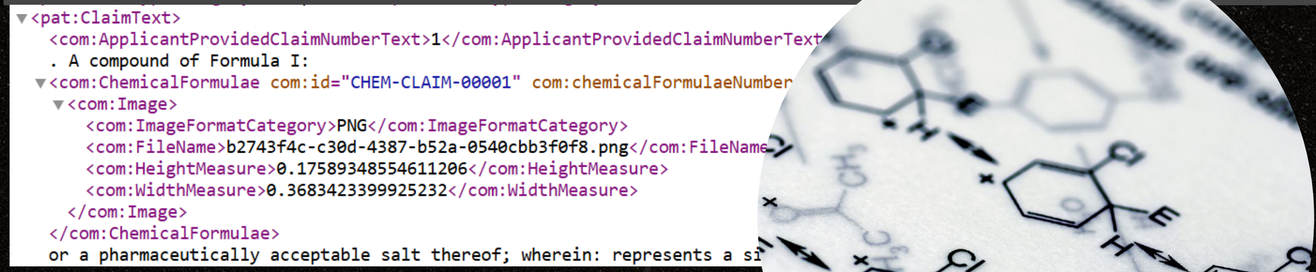


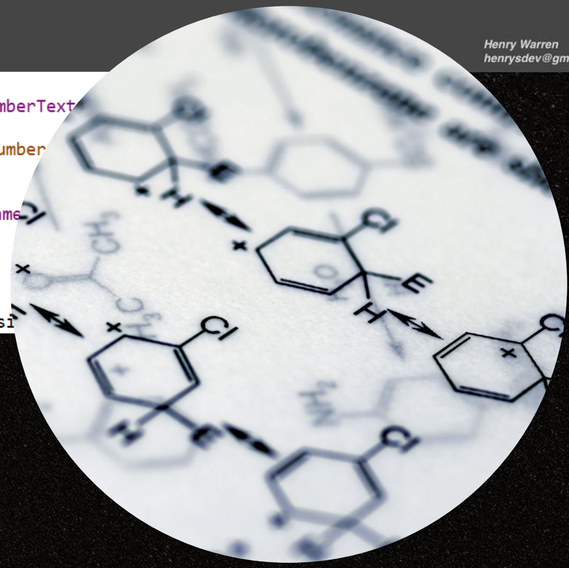
P TM D PBR

Chemical image classification

- Convolutional neural network trained for an image classification task
- Trained on 20,000 images extracted using Textract from approx. 2000 Australian patents







99.7%

accuracy identifying an image containing a chemical formula

Math formula detection

- Recognising math formulas is very challenging
- Formulas look like text
- Often embedded within text (inline)
- Need to know where a formula begins and ends
- Textract does not find formulas as figures (unlike chem)
- No state-of-the-art tools

The answer to Gauss equation, the $(x - a_1)/b_1 \equiv 118286 \bmod 337 * 577 \equiv \sqrt{1}$ is given by the root obtained by the Pythagorean Triples method multiplied by the $\sqrt{1} \bmod m$.

Thus part of the root is obtained by this method, but not the whole root needed.

Two other anchors, besides $\sqrt{1} \bmod p * q$, for the square and root equations are possible:

1. $2 * (a + b_1^2 + b * \sqrt{1}) \bmod p * q \equiv (a_1 + b_1 * \sqrt{1})^2$
2. $2 * (a + b * \sqrt{-1}) \bmod p * q \equiv (a_1 + b_1 * \sqrt{-1})^2$
3. $(2 * a + (2 * b + b_1^2) * \sqrt[3]{-1}) \bmod p * q \equiv (a_1 + b_1 * \sqrt[3]{-1})^2$

Feeding in the following parameters will return natural squares for when the anchor is $\sqrt{-1} \bmod p * q$. The actual square root of -1 will be found when the square is found. The equation that has to be solved in order to find the first natural square is:

$2 * (lp1^2 + lp2^2) + 1 \bmod p * q \equiv 2 * (lp1 + lp2)$

There seems to be no easy way to solve this equation. The output for the procedure for the first 10 natural squares follows:

Math formula detection (cont'd)

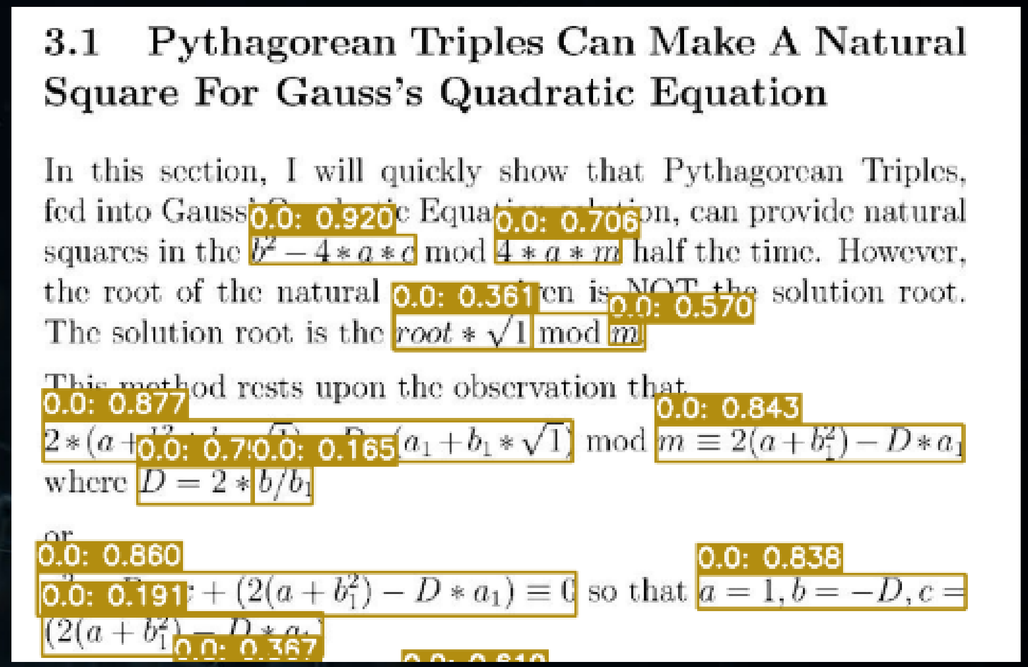
- Image recognition CNN
- Trained on **FormulaNet** dataset
- Around 2 million labelled formulas across ~47k pages

Accuracy

96.7%
standalone formulas

92%
inline formulas

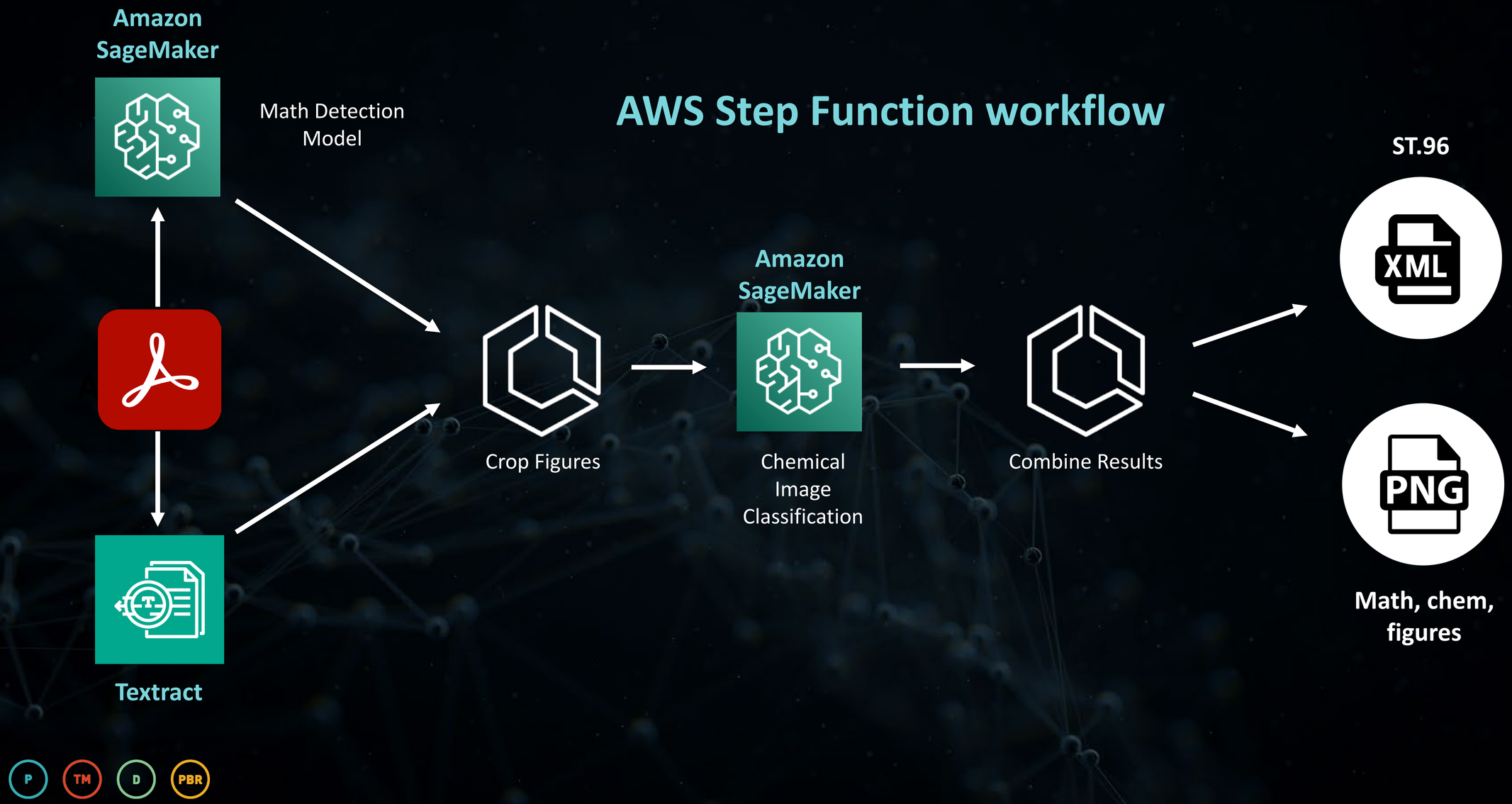
94%
OVERALL




```
▼<com:P com:id="p-00001" com:pNumber="1">
  Paul C. Bingham Cheffers 21 June 2018 Using the Gauss Modular Quadratic Equation, and using modular complex numbers mod
  ▼<com:Math com:id="MATH-ABSTRACT-00001" com:mathNumber="00001">
    ▼<com:Image>
      <com:ImageFormatCategory>PNG</com:ImageFormatCategory>
      <com:FileName>58b909db-9d6d-474e-bf69-b315fdcb3c5f.png</com:FileName>
      <com:HeightMeasure>0.012924879789352417</com:HeightMeasure>
      <com:WidthMeasure>0.0285519063472747 8</com:WidthMeasure>
    </com:Image>
  </com:Math>
  a modular quadratic equation:
  ▼<com:Math com:id="MATH-ABSTRACT-00002" com:mathNumber="00002">
    ▼<com:Image>
      <com:ImageFormatCategory>PNG</com:ImageFormatCategory>
```


P TM D PBR

ST.96 generation flow



Beyond images

- Classification and recognition tasks can also be applied to text
- Classifying paragraphs to identify description sections such as BackgroundArt, EmbodimentDescription and TechnicalField
- Named Entity Recognition to identify and mark PatentCitation, NPLCitation, BioDeposit etc

Thank you



Australian Government
IP Australia

Contact us

International.ICTCooperation@ipaustralia.gov.au

ipaustralia.gov.au

facebook.com/ipaustralia.gov.au

x.com/IPAustralia

linkedin.com/company/ip-australia

youtube.com/user/ipaustralia

instagram.com/ipaustraliaofficial/

P TM D PBR

Chemical formula types

A guide to formulae in organic chemistry

COMPOUND INTEREST

Molecular formula

The molecular formula of an organic compound simply shows the number of each type of atom present. It tells you nothing about the bonding within the compound.

$C_4H_8O_2$

Molecular formula of butanoic acid

Empirical formula

The empirical formula of an organic compound gives the simplest possible whole number ratio of the different types of atom within the compound.

C_2H_4O

Empirical formula of butanoic acid

Condensed formula

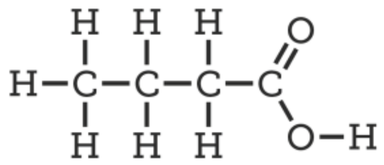
In condensed formulae, each carbon atom is listed separately, with atoms attached to it following. In cyclic parts of molecules, like benzene, carbons are grouped.

$CH_3CH_2CH_2COOH$

Condensed formula of butanoic acid

Displayed formula

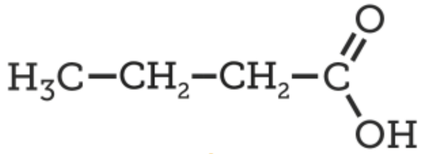
A displayed formula shows all of the atoms and all of the bonds present in an organic compound. The bonds are represented as lines.



Displayed formula of butanoic acid

Structural formula

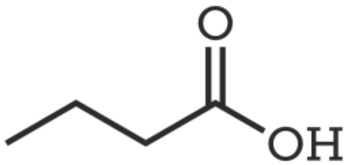
Similar to displayed formula; not all bonds are shown, although all atoms are still indicated using subscript numbers. Carbon hydrogen bonds are often simplified.



Structural formula of butanoic acid

Skeletal formula

In skeletal formulae, most hydrogen atoms are omitted. Line ends or vertices represent carbons. Functional groups and atoms other than carbon or hydrogen are still shown.



Skeletal formula of butanoic acid

P TM D PBR