JISC The Joint Information Systems Committee

EPSRC

E-Malaria: A computer-aided drug discovery system for chemistry teaching

Jeremy Frey

School of Chemistry
University of Southampton

Malaria

- Malaria kills over 2 million annually
- Caused by a parasite and transmitted by certain mosquitoes
- New drugs are needed as resistance to existing drugs grows
- Computational modeling can speed up process and save laboratory time and costs
- Relevant project to all students from High School to University

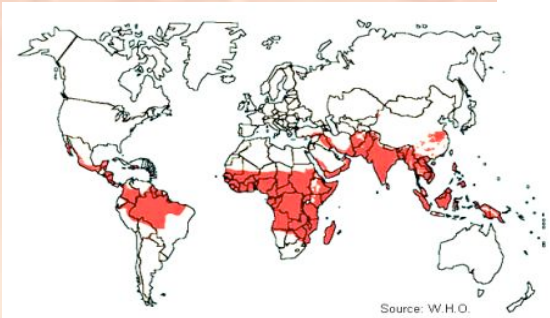


Figure W.H.O. map of chloroquine resistance in *P. falciparum*.

Initially a School outreach project
Why target school pupils?

- Numbers of pupils choosing science courses are falling
- Science is perceived as boring, hard and irrelevant to peoples lives
- Decline is numbers is worrying for the science community and society at large

What difference can the e-Malaria project make?

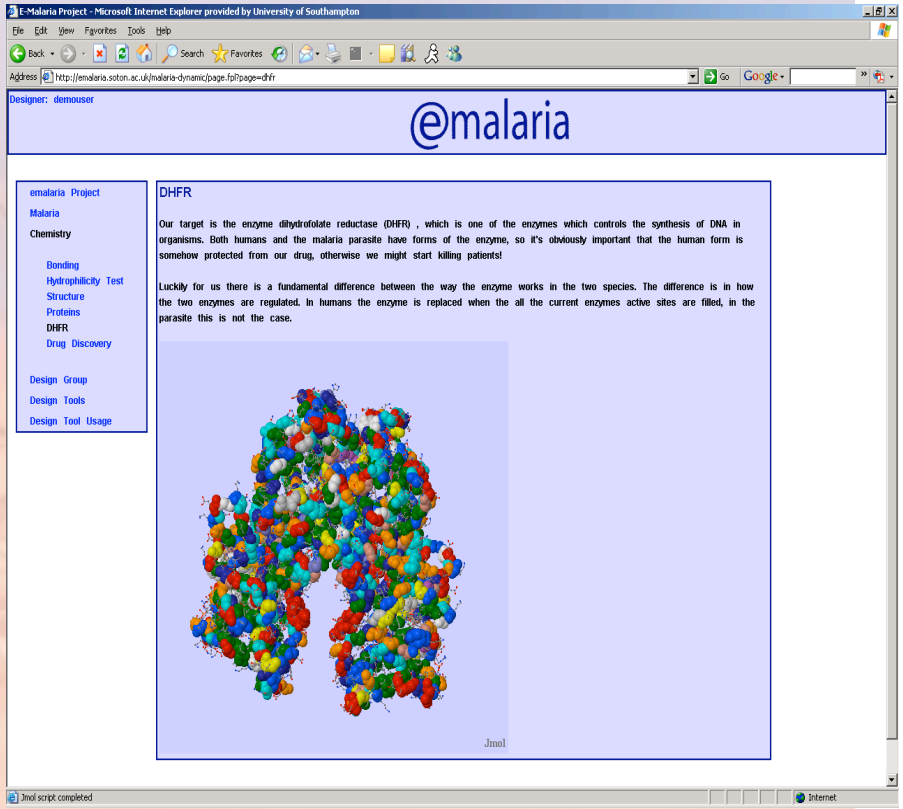
- Example of chemistry in context
- Authentic activity
- Students use real data and real software
- Chance drug candidates could go on for *in vitro* & *in vivo* tests

What does the project provide?

- Range of supporting texts and links
- Interactive quizzes
- Forum
- Links with university departments
- Support from project team
- Drug design tools

Resource design

- Based on feedback from range of project supporters
- Designed to look contemporary and interesting
- Accessibility for students with SEN (Special Educational Needs)
- Interactivity for interest



Interactivity

- Core to keeping students involved
- Increases the amount learnt, understood and memorized by students
- Provides interest
- Develop skills in more realistic and relevant situations

Supporting Material

- Intermolecular forces
- Drug design
- Proteins and amino acids
- Enzymes
- Also linked to material on statistics, experimental design and wider chemical informatics

Design

- Take a suitable enzyme target in the malaria parasite (DHFR)
- Design small molecule as possible drug
- 'Dock' in to enzyme target to find improved binding
- Modify to yield drug like molecule

Current user:
WWW
LOGIN

@malaria

emalaria Project
- Background
- Southampton
- Schools
- Escience
- CCDC
- Project Statistics

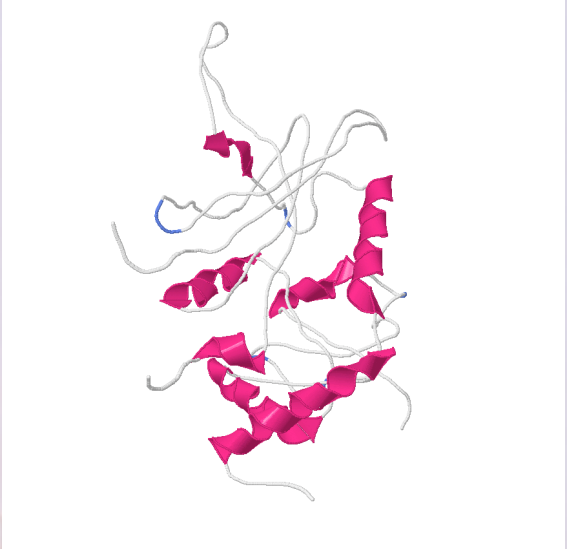
Malaria

Chemistry

Welcome to the e-malaria Project!

The Target - DHFR

- Regulates part of DNA synthesis
- Present in both humans and parasites
- Different regulation methods between humans and parasites make it an excellent target



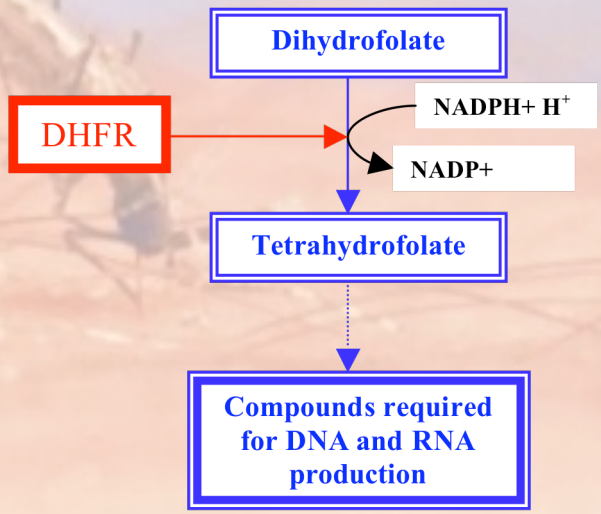
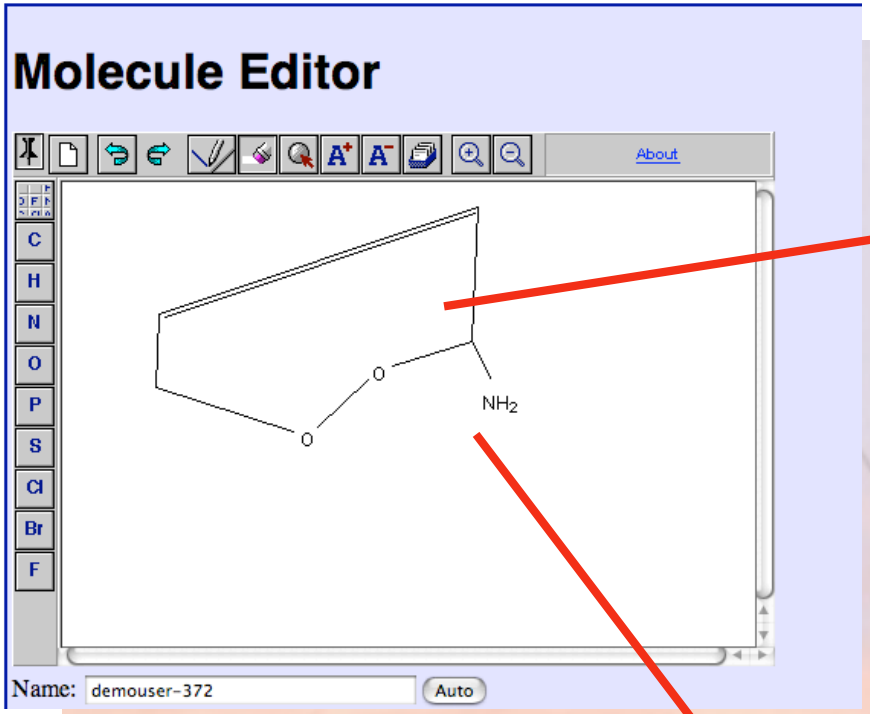


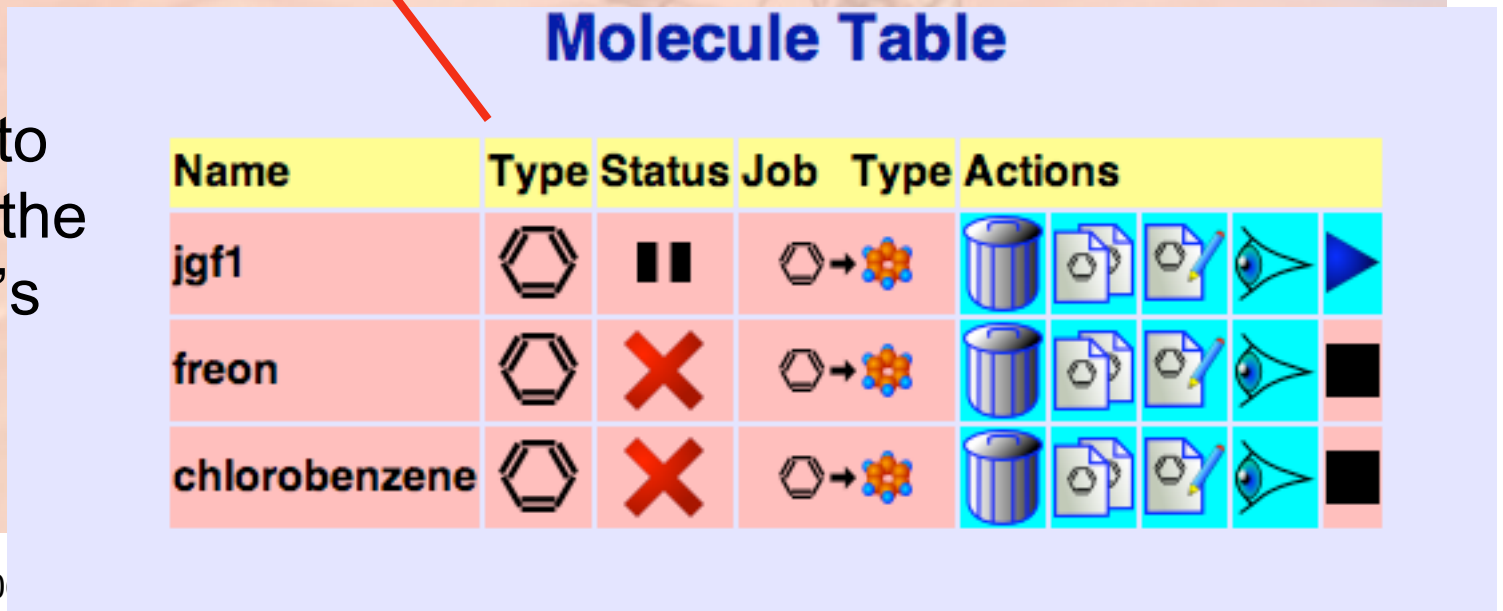
Figure 0 Normal Folate reaction Pathway



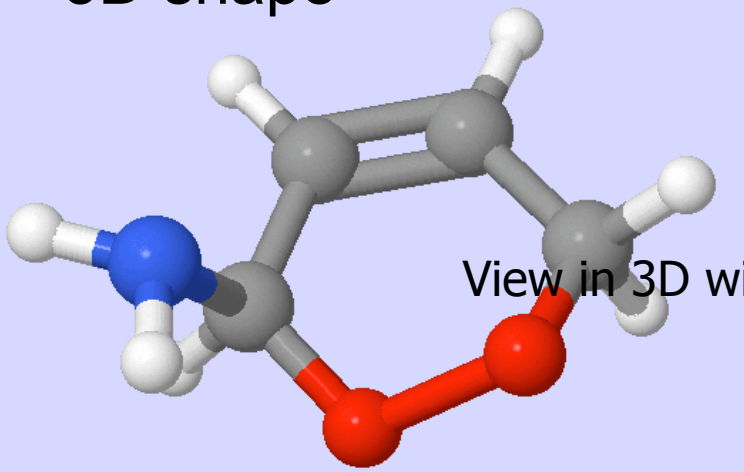
Design a possible drug

Convert to a reasonable 3D shape

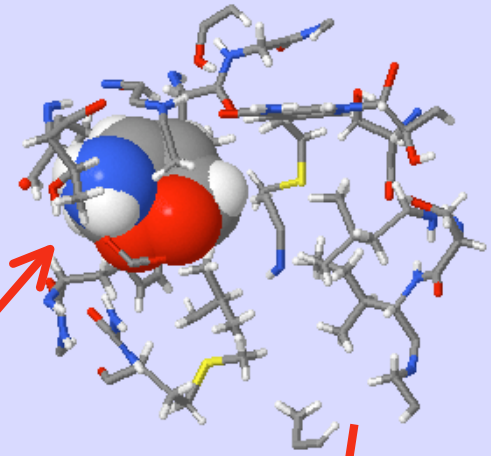
Use QM program to work out the molecule's shape



3D shape



View in 3D with Jmol



Now try to dock the drug in to the enzyme active site- but which way round? Lots of ways to try!

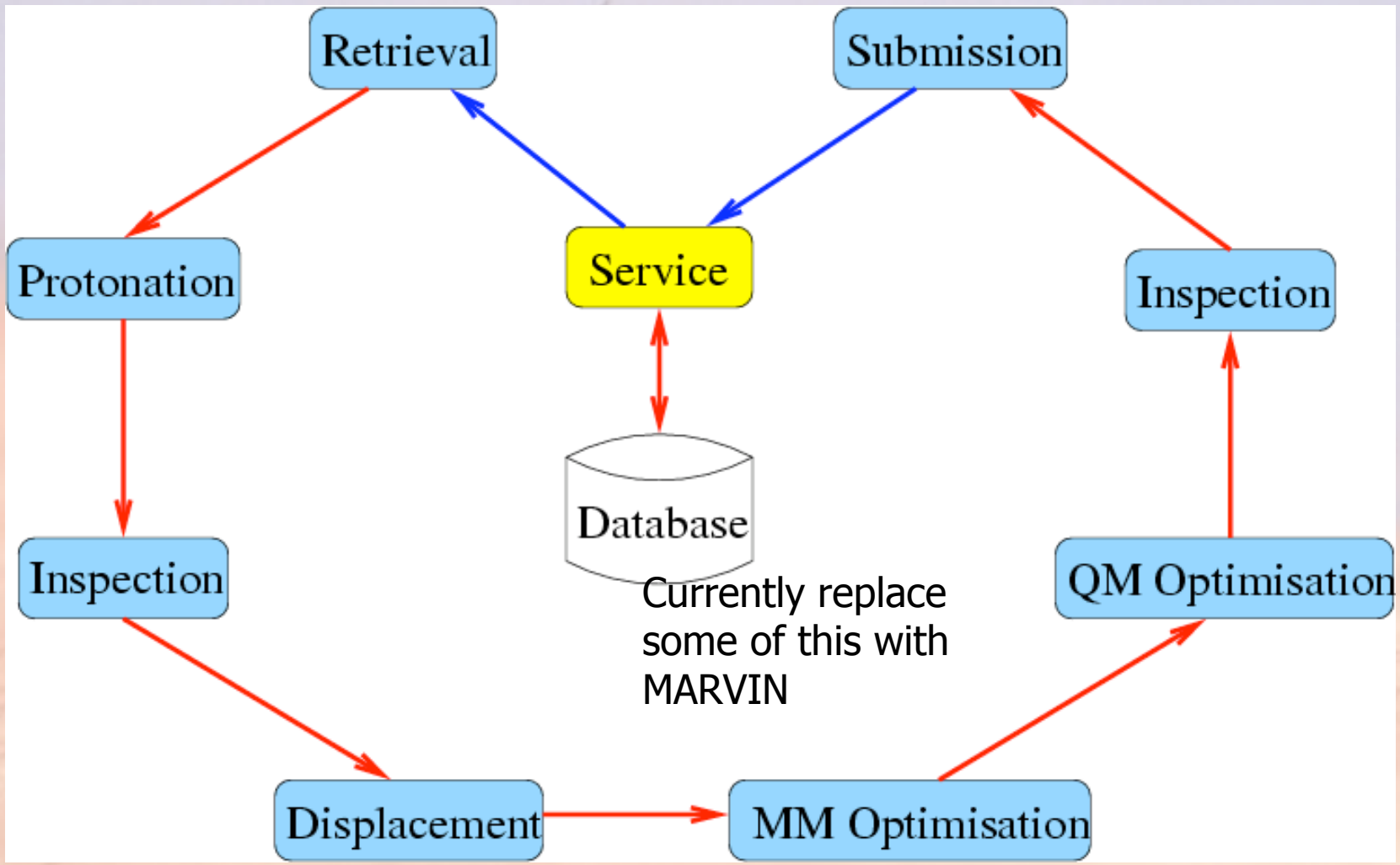
How well does it bind?

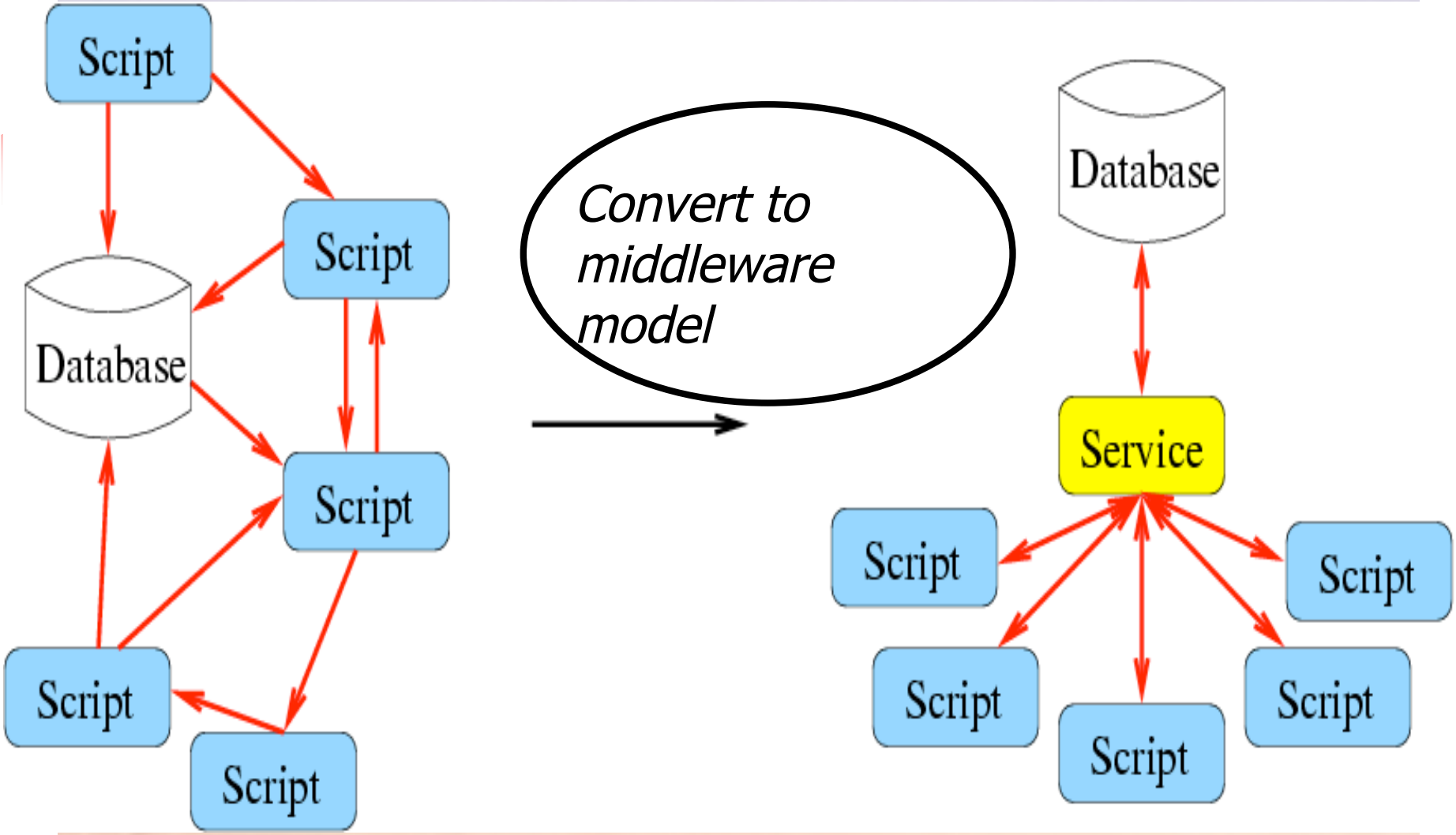
Molecule Table

Name	Type	Status	Job Type	Actions
jgf1				
freon				
chlorobenzene				

Results Table

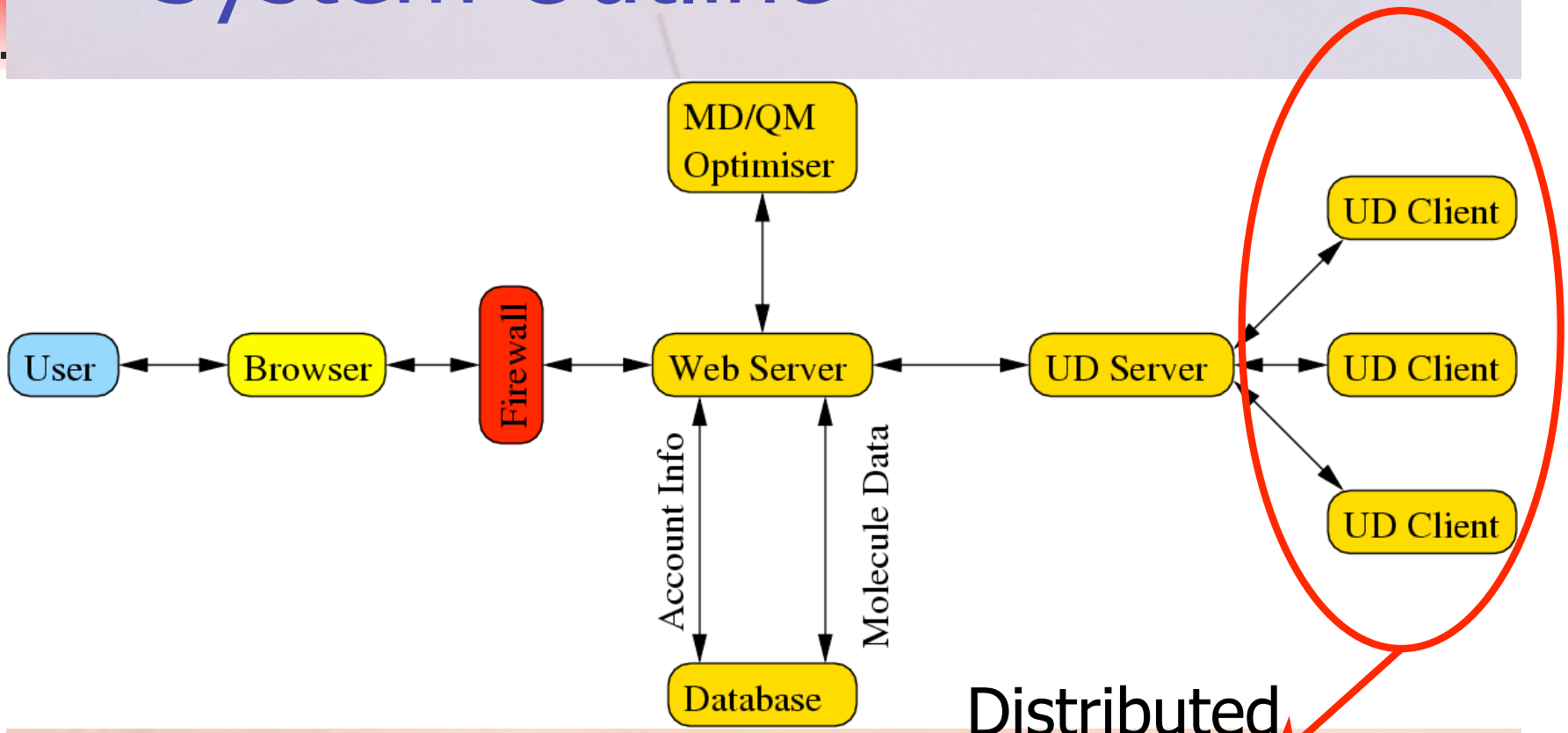
Name	H-bond	VdW	Strain	Energy	
jgf1	13.46	15.97	-1.16	34.25	
benzene	0	17.54	0	24.12	
demouser-362	0	11.3	-0.64	14.9	


Retrieval
Submission
Protonation
Service
Inspection
Database
Inspection
QM Optimisation
Currently replace some of this with MARVIN
Displacement
MM Optimisation



Molecular Structure File Format Conversion

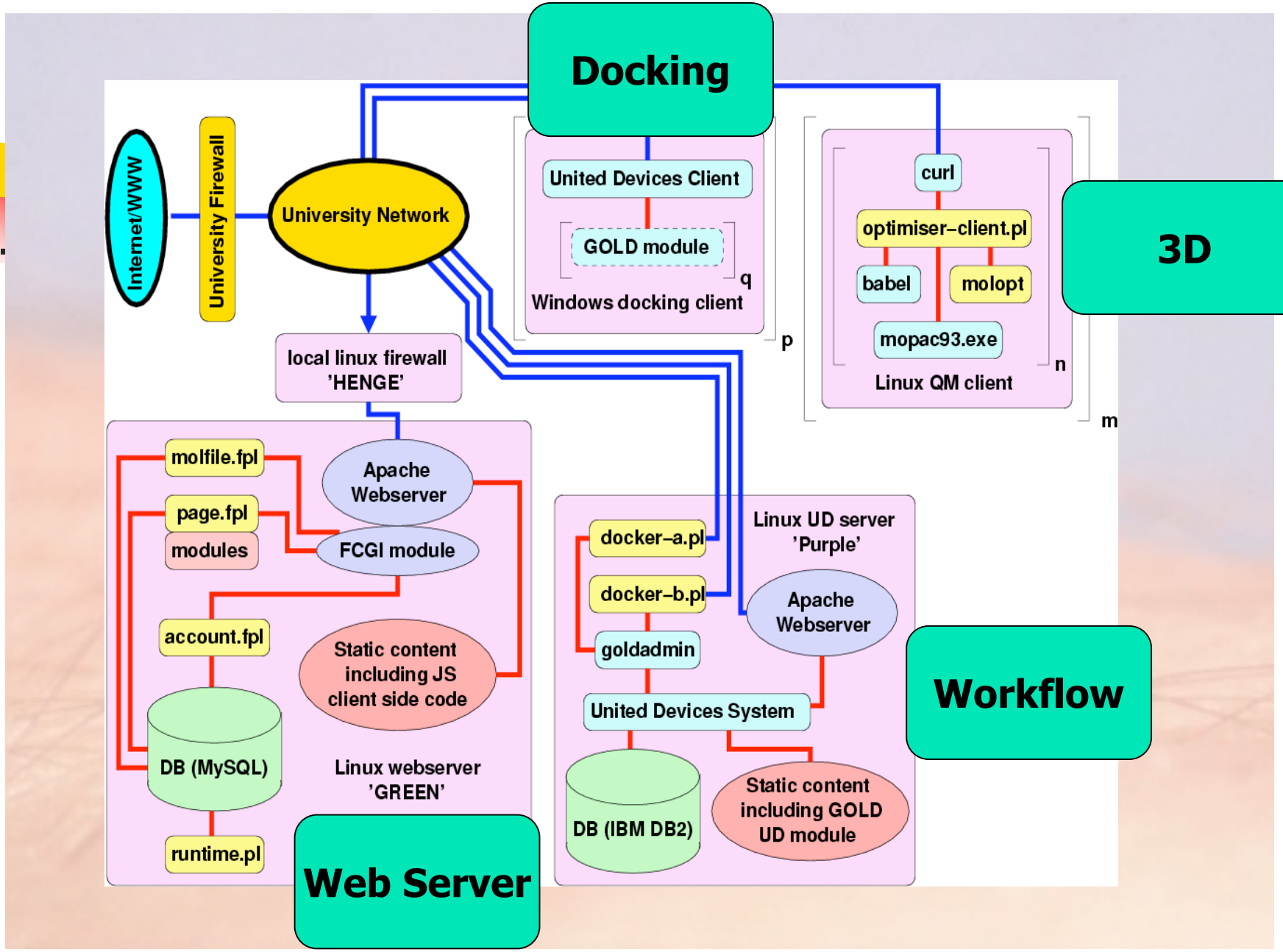
System Outline



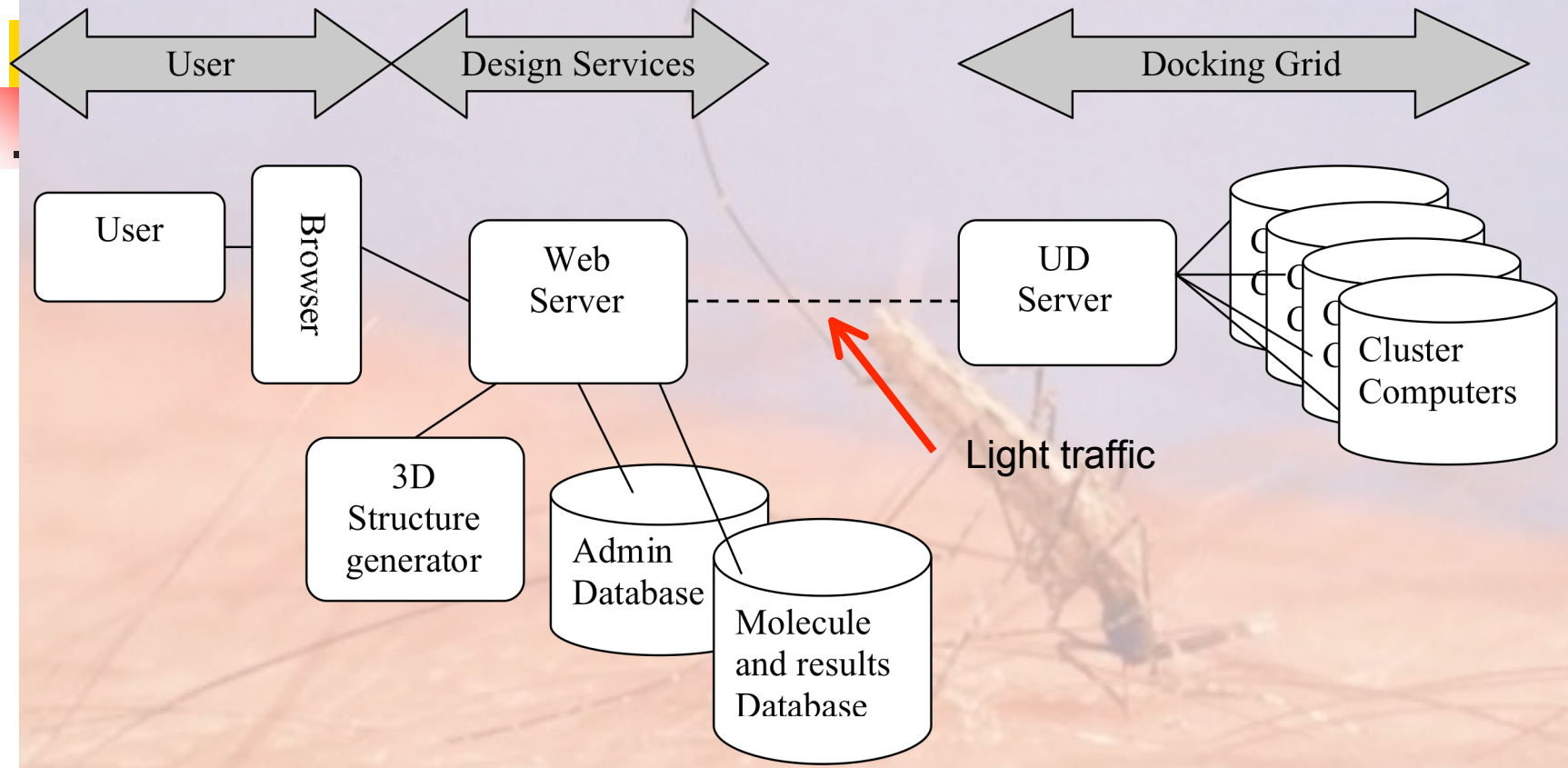
Distributed computing cycle steeling grid

Why UD

- UD software is relatively heavy weight but highly secure
- Need to be secure to allow us to run the GOLD software (from CCDC)
- This is real and valuable software which must be protected.
- Don’t have to worry about invalid answers as we can always readily check


Docking
Internet/WWW
University Firewall
University Network
United Devices Client
GOLD module
q
Windows docking client
p
curl
optimiser–client.pl
babel
molopt
mopac93.exe
n
Linux QM client
m
3D
local linux firewall
'HENGE'
molfile.fpl
page.fpl
modules
Apache
Webserver
FCGI module
account.fpl
Static content
including JS
client side code
DB (MySQL)
Linux webserver
'GREEN'
runtime.pl
Web Server
docker–a.pl
docker–b.pl
goldadmin
Linux UD server
'Purple'
Apache
Webserver
United Devices System
DB (IBM DB2)
Static content
including GOLD
UD module
Workflow

The functional outline of e-Malaria project



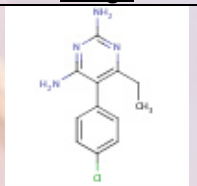
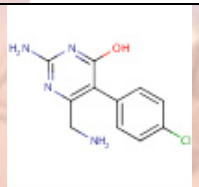
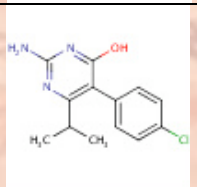
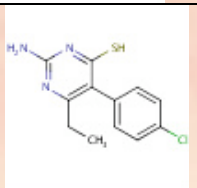
User interaction via a browser with the web server that coordinates the sequence of calculations to convert a 2D rough sketch of a molecule to a reasonable 3D structure using dedicated resources and then performs a docking calculation against the malarial DHFR protein using the UD cycle steeling clusters.

Current drug

- Trimethoprim,
- score 57.49

Table 1 Individual Job list

Job Name	SMILES	H-bonding	VdW	Strain	Energy
Pyrimethamine	CCC1=C(C(N)=NC(N)=N1)C2=CC=C(Cl)C=C2	4.09	40.18	-0.97	58.36
Pyrimethamine 2	NCC1=C(C(O)=NC(N)=N1)C2=CC=C(Cl)C=C2	9.68	39.58	-3.31	60.79
Pyrimethamine 3	CC(C)C1=C(C(O)=NC(N)=N1)C2=CC=C(Cl)C=C2	5.06	41.80	-1.29	61.24
Pyrimethamine 4	CCC1=C(C(S)=NC(N)=N1)C2=CC=C(Cl)C=C2	6.62	41.59	-1.62	62.18

Molecule	Image
Pyrimethamine	
Pyrimethamine 2	
Pyrimethamine 3	
Pyrimethamine 4	

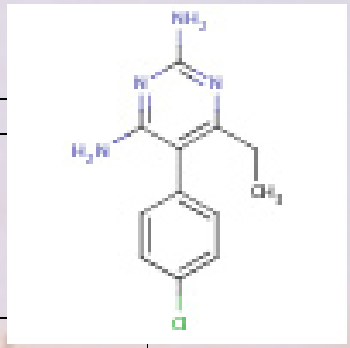
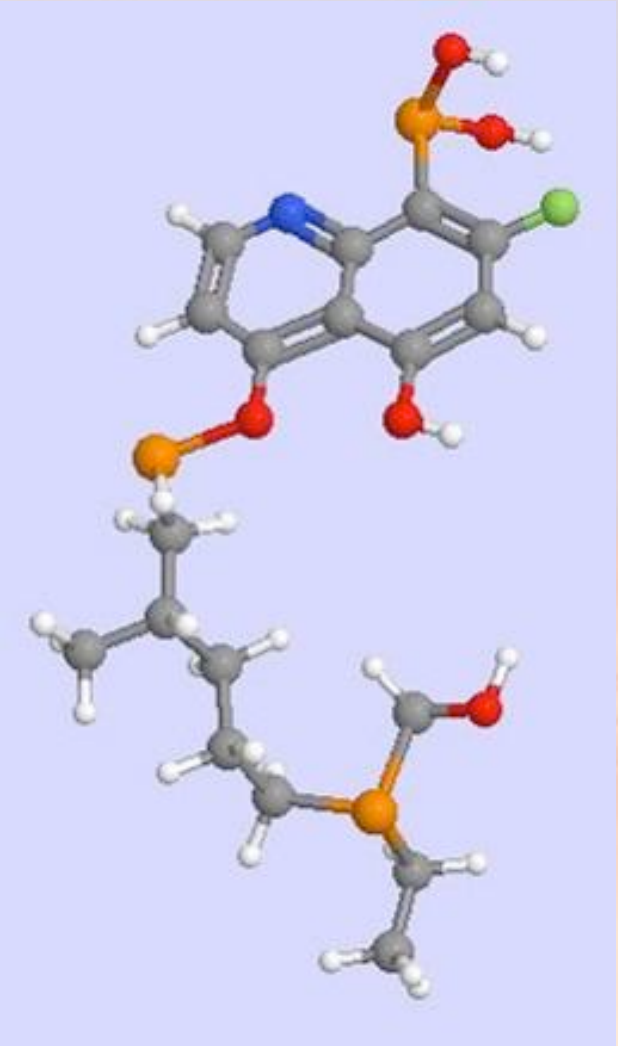


Table 2 Molecular structures

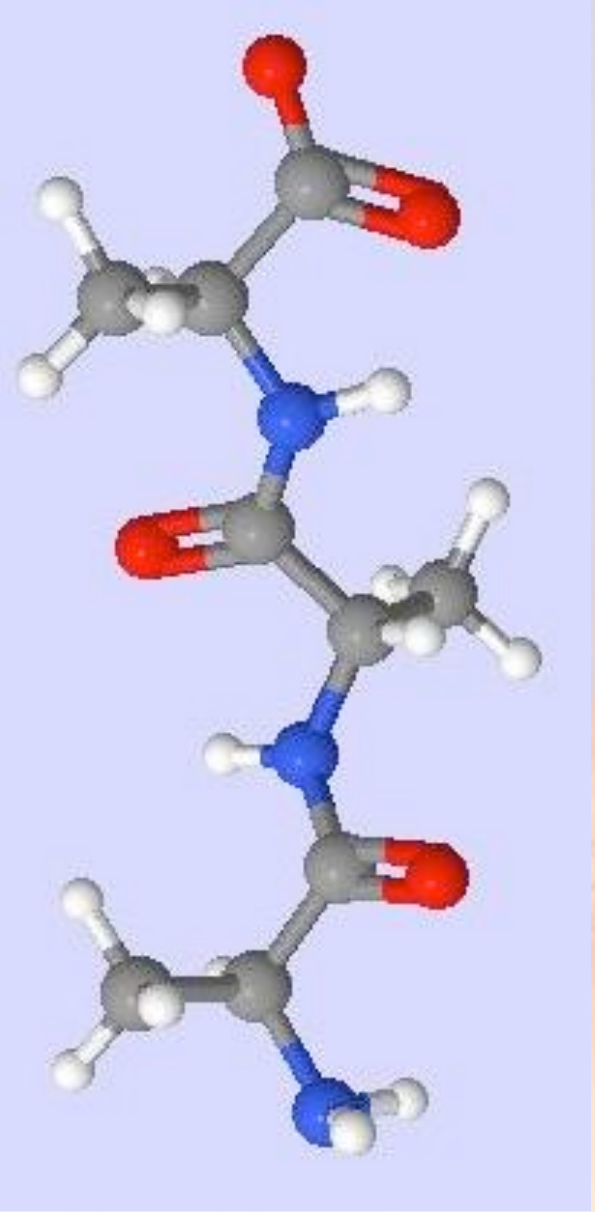
Organo-phosphorous?

- One student found a high scoring compound
- score 68.1
- Raised the question of what else does it bind to?
- We suggested they look up information on nerve gases?

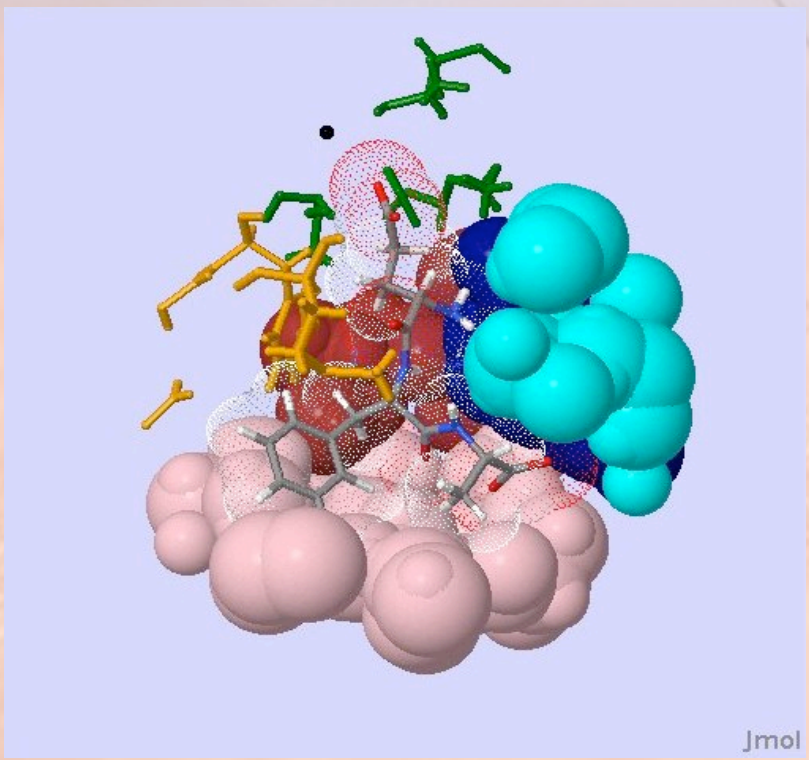


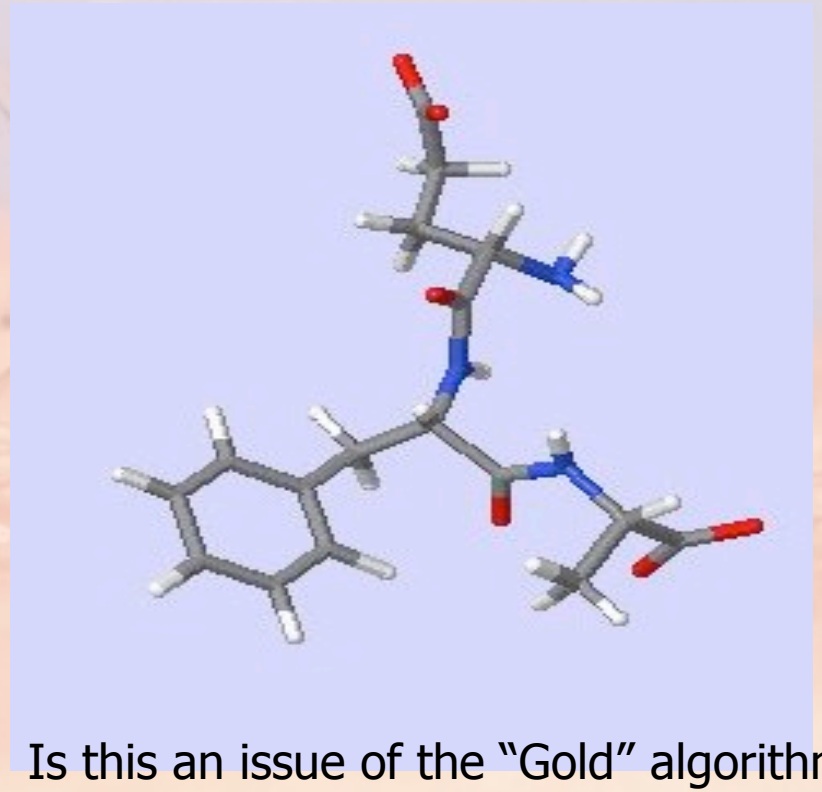
Peptides as drugs?

- **Ala-ala-ala (tripeptide)**
- **score 50.15**
- **Suitable as a drug?**
- **Can build these relatively easily**
- **Screen many peptides**



Docked conformation of Glu-Phe-Ala, score 68.88 surprisingly large value!





Is this an issue of the "Gold" algorithm?

Chemical Informatics (Year 4 Course)

Making use of the e-Malaria system

- A progressive building up of complexity and use of real and authentic data, clearly linked and of relevance to work-based learning,
- Timely as the students had just completed their six-month work placement.
- Active engagement with datasets enabled students to understand difficult concepts for the first time
  - by being able to experiment and use
  - for example linear regression
  - use in context enabled them to be able to see the relevance and value of such techniques.

CHEM 6016 Mini-Project Part B

CHEM 6016 Mini-Project Part B (Blackboard under Assignments)

From the National Cancer Institute (NCI) CACTUS data bases (go to cactus.nci.nih.gov and select the enhanced NCI Database Browser) perform a Query by selecting 'Random Set' from the drop down menu in the first Query Type Box. I suggest you ask for 50 compounds to be returned (box on the lower right hand side).

From the set of molecules returned, select out compounds with molecular weights between 200 and 400. Remove any ions and any molecules containing elements other than C, H, N, O, S, P, F, and Cl and reduce your set to 12 molecules.

For each of the molecules in your set of 12

1. Use the e-Malaria system to draw in these molecules and obtain 3D structures.
2. Compare the 3D structures obtained in this way with the ones in the CACTUS database (if present) and any crystal structures of the molecules that you can find.
3. Run the docking program and obtain a docking score for the molecule with the Malarial DHFR protein. You will need to login to the emalaria web site (emalaria.soton.ac.uk) using the username and password supplied. Draw your molecules in 2D using the applet provided and use the system to convert them to 3D (check the stereochemistry) and then dock the molecule in to malarial DHFR. You may want to run the docking calculation more than once for each molecule. The user guide for eMalaria is provided under Course Documents on Blackboard.
4. Derive values for a set of descriptors for the molecule. Some of these, such as molecular weight, no. H bond acceptors, etc are provided by CACTUS but others such as LogP, surface area, will need to be calculated. One web site that allows you to calculate descriptors can be found at http://www.molinspiration.com/cgi-bin/properties. You may be able to find other sites which calculate other descriptors. You may also find experimental values for physical properties on for example the PhysProp databases accessed via CDS. You should aim for between 5 and 10 descriptors for each molecule.
5. Explain why you have chosen these descriptors.

For 10 of these molecules

6. Using ideas from the course and experience from Part A, build a QSAR model to predict the e-Malaria docking score. Explain how you build the model including the selection of the most appropriate descriptors. Justify your selection and the nature of the model you have built.

Using the two remaining molecules

7. Test the predictive ability of the model with the two molecules not used in constructing the model.
8. Discuss how the model could be improved.

You should write up this section of the mini-project as a short report, to journal standard, along with the report from the workshop on regression (Mini-Project Part A). You should present a summary of your report as a web page. [Saving a WORD file as a 'Web Page, or HTML' will generate a suitable web page from the WORD document, and the University web site has details of how to put your file on your University account and make it publicly available]. The presentations we will ask you to give should primarily cover the work of this part of the Project (Part B) but may certainly refer to experience gained from Part A.

NCI Database

Feed 12 molecules in to e-Malaria

Look at calculated 3D structures and check stereochemistry

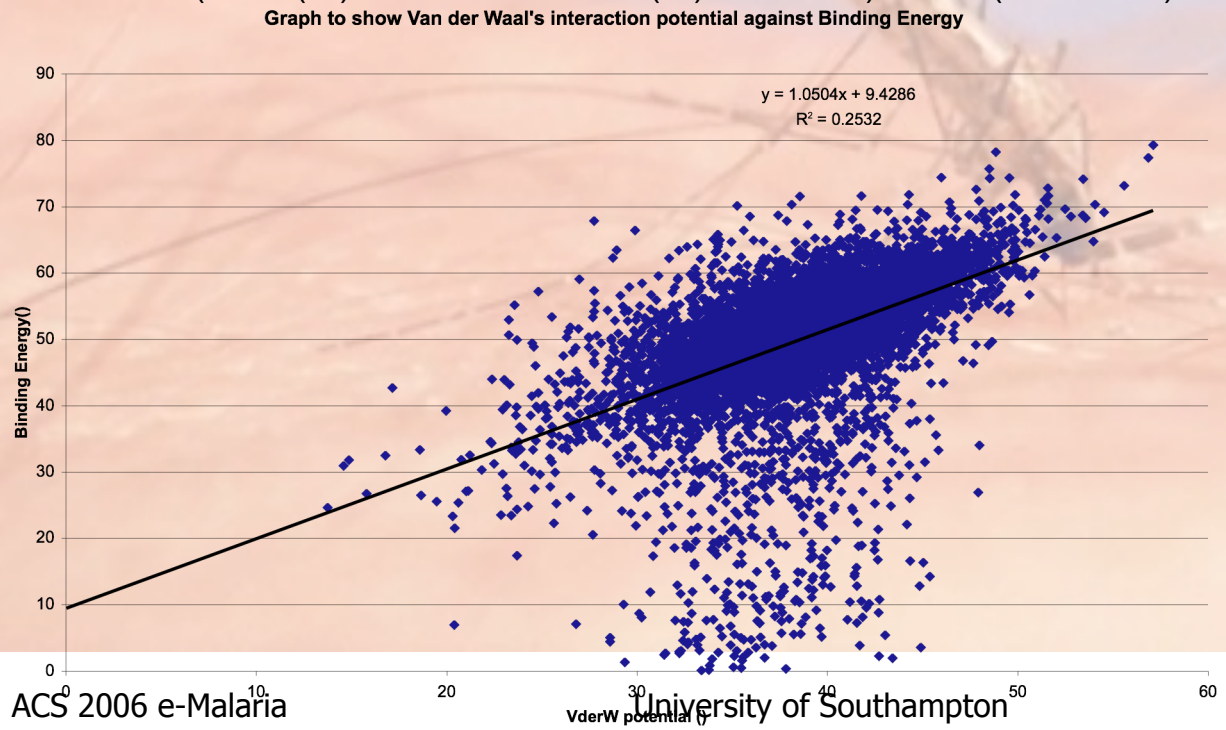
Chose & calculate descriptors

Build, test and evaluate statistical models

Write a presentation

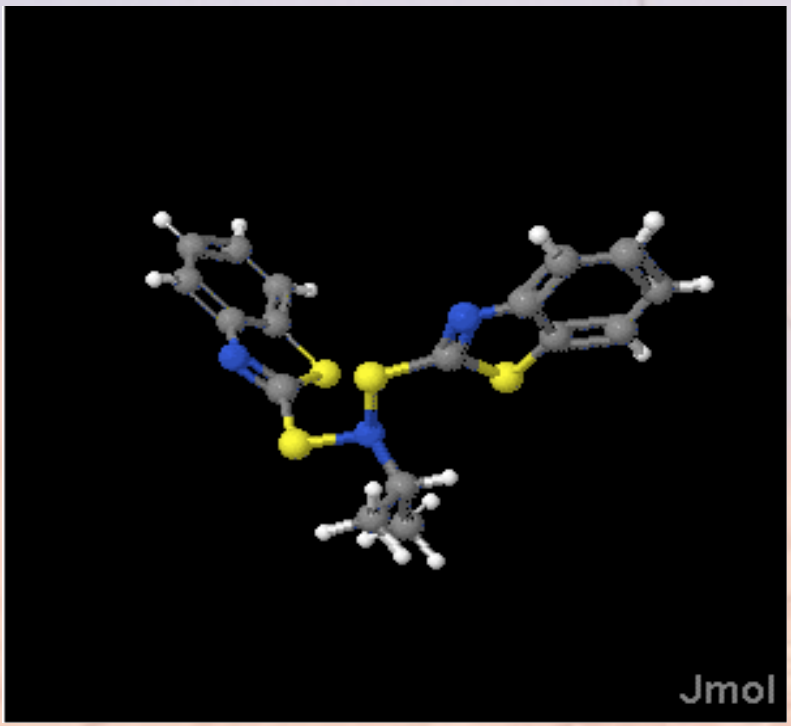
NCI database screen

	CAS	Smile string	Gold Score
C17H15N3	3741-79-5	CC(N(SC2=NC1=CC=CC=C1S2)SC4=NC3=CC=CC=C3S4)C	79.31
C18H18N4	5460-88-8	C(C2=NC1=CC=CC=C1[NH]2)CCCC4=NC3=CC=CC=C3[NH]4	78.27
C18H11Cl5	5273-31-4	S(C1=C(C(=NN=C1)Cl)SCC2=CC(=C(C=C2)Cl)Cl)CC3=CC(=C(Cl)C=C3)Cl	77.38
C16H14O4	5325-76-8	O=C(CC1=CC=C(C=C1)SSC2=CC=C(CC(=O)O)C=C2)O	75.74
C14H10N4	1155-37-9	C2(=NC1=CC=CC=C1[NH]2)SSC4=NC3=CC=CC=C3[NH]4	74.44
C16H12N2	6949-41-3	C3(=CC=C(N=NSC2=CC1=C(C=CC=C1)C=C2)C=C3)S(=O)(=O)O	74.39
C12H14N4	3905-92-8	O=S(N)(=O)C1=CC(=C(C=C1)SSC2=C(C=C(C=C2)S(N)(=O)=O)N)N	74.33
C17H14N4	95-35-2	C(NC(=O)NCSC2=NC1=CC=CC=C1S2)SC4=NC3=CC=CC=C3S4	74.2
C18H15Cl2	6299-26-9	C(C1=C(Cl)C=CC=C1)SC2=NC(=NC(=C2)N)SCC3=C(C=CC=C3)Cl	73.2
C18H13Cl3	5273-30-3	S(C1=C(Cl)N=NC=C1SCC2=C(Cl)C=CC=C2)CC3=C(C=CC=C3)Cl	72.82



Score vs van der Walls energy

Best "Gold" score from NCI database



(((1,3-benzothiazol-2-ylthio)(isopropyl)amino)thio)-1,3-benzothiazole

Interactive & relevant

- Access to data and ‘real’ results, enabled students to interact and hence develop their own understanding
  - Before the course I hadn’t really considered how the computer actually does it but … interesting to see how that works and then there is a part of the course where .. they taught you… how to interpret data and build models .. and that’s probably quite a useful part of that project use the model building … its all very well people telling you this is a …peptide but until you actually use it, you cant really visualise it. [Int 5]
- Industrial speakers to put the work in context

Project Work

- Di- and Tri-peptides as possible drugs
- Use the power of the computational grid to screen all the possible combinations of the 21 natural aminoacids.
- Develop automated ways of "constructing the peptides"
  - Smile Strings to the rescue
- Exploit and develop "chemical knowledge" to interpret and understand the resulting mass of data

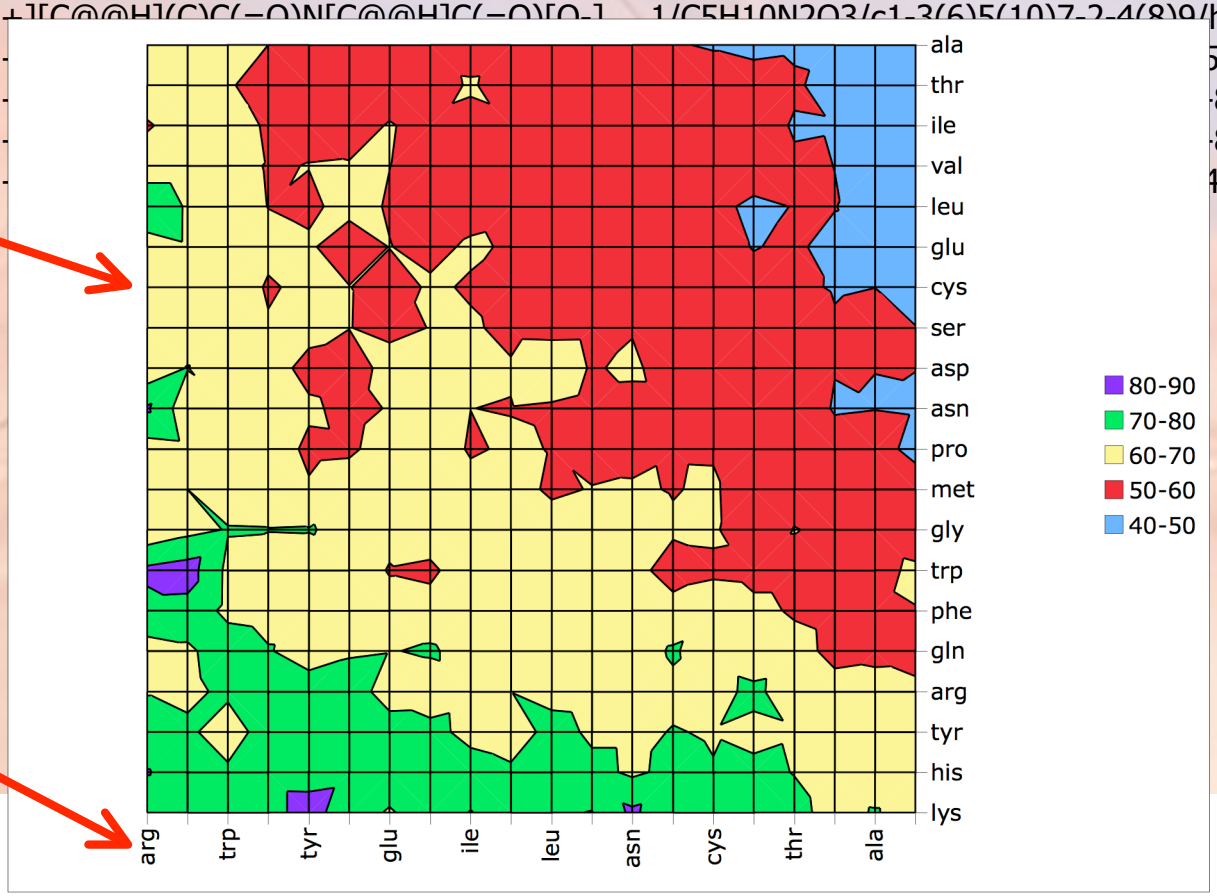
Di-peptides

Smile string

Min score

Max score

ala-ala	[NH3+][C@@H](C)C(=O)N[C@@H](C)C(=O)[O-	1/C6H12N2O3/c1-3(7)5(9)8-4(2)6(10)11/h3-4H,7	10	38	46
ala-arg	[NH3+][C@@H](C)C(=O)N[C@@H](CCCNC(=N)N	1/C9H19N5O3/c1-5(10)7(15)14-6(8(16)17)3-2-4-	10	53	64
ala-asn	[NH3+][C@@H](C)C(=O)N[C@@H](CC(=O)N)C(	1/C7H13N3O4/c1-3(8)6(12)10-4(7(13)14)2-5(9)1	10	46	52
ala-asp	[NH3+][C@@H](C)C(=O)N[C@@H](CC(=O)[O-]	1/C7H12N2O5/c1-3(8)6(12)9-4(7(13)14)2-5(10)1	10	47	51
ala-cys	[NH3+][C@@H](C)C(=O)N[C@@H](CS)C(=O)[O	1/C6H12N2O3S/c1-3(7)5(9)8-4(2-12)6(10)11/h3-	10	44	49
ala-gln	[NH3+][C@@H](C)C(=O)N[C@@H](CCC(=O)N)C	1/C8H15N3O4/c1-4(9)7(13)11-5(8(14)15)2-3-6(1	10	48	54
ala-glu	[NH3+][C@@H](C)C(=O)N[C@@H](CCC(=O)[O-	1/C8H14N2O5/c1-4(9)7(13)10-5(8(14)15)2-3-6(1	10	46	55
ala-gly	[NH3+][C@@H](C)C(=O)N[C@@H]C(=O)[O-]	1/C5H10N2O3/c1-3(6)5(10)7-2-4(8)9/h3H,2,6H2,	10	39	44
ala-his	[NH3-	5)16)2-6-3-	10	50	60
ala-ile	[NH3-	8(12)6(3)1	10	46	53
ala-leu	[NH3-	8(12)6(3)1	10	44	53
ala-lys	[NH3-	4)15)4-2-3-	10	51	62



Future - Extending Design Ideas

- Use to teach statistical design of experiments (DoE)
- Use DoE to make a informed choice of the subset of the the NCI database based on suitable ranges of descriptor values
- Attempt to make a rational coverage of "Chemical Space"

Issues

- Providing instructions to a wide range of users
- Competition in the schools
  - Need to provide personal, school and overall summaries
- Keeping the systems running
  - Robust web server & software
  - Differences between browsers!
  - Log file overload
  - Network problems

frustrations

- The inevitable technical problems, which led to frustrations on the part of the students.
  - [E-Malaria site] I did use it but it was quite frustrating… I think its licence ran out… but problem was I didn't realise anything was wrong with it, it hadn't been mentioned… I might have been the first person to try and do it… loaded all my molecules in… done half of it that was the thing … I didn't realise anything was wrong.. and then it wouldn't work out …. Which was the important bit… and so that was frustrating. [Int 3]

Related projects

- Related to other seti@Home projects
  - Graham Richards drug screen
  - Climate prediction
- But student designs molecules not just supply computer power to screen someone else's choice of a possible drug
- Student sees and plays with input & output
- More complex exchanges between us and the students, but data volumes not large, but frequent

People & Organizations

- Rob Gledhill
- Sarah Kent
- Andrew Milstead
- Brian Hudson
- John Metcalfe
- John Frampton
- Havant College
- Jon Essex
- Graham Richards
- CCDC
- UD
- JISC
- EPSRC