NMRbox: Toward Reproducible Computation for Bio-NMR

Jeffrey C. Hoch, Ph.D.

UConn Health

Overview

- What is NMR?
- Biomedical applications
- Computation in NMR
- Challenges, especially reproducibility
- Application of virtualization and clouds
- Challenges awaiting solutions

The mission of the Center is to

- Simplify the development, discovery, maintenance, and use of bio-NMR software
- Facilitate optimal and reproducible workflows
- Provide access to advanced training and computational resources
- Serve as a test bed for applying emerging computing technologies to bio-NMR
- *Foster reproducibility of bio-NMR analyses*

People

UConn

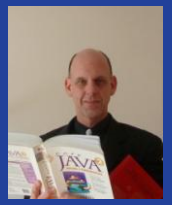
Michael Gryk

Mark Maciejewski

Adam Schuyler

Ion Moraru

Oksana Gorbatyuk



Gerard Weatherby

Dilllon Jones

Wisconsin

Eldon Ulrich

Miron Livny

Hamid Eghbalnia

Dmitry Maziuk

Vincent Chen

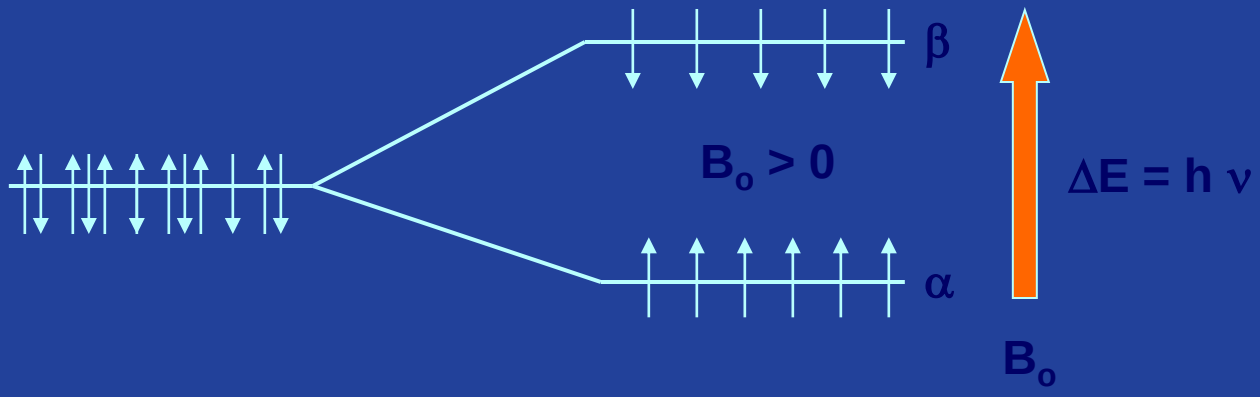
Jonathan Wedell

Pedro Romero

Kumaran Baskaran

What is NMR?

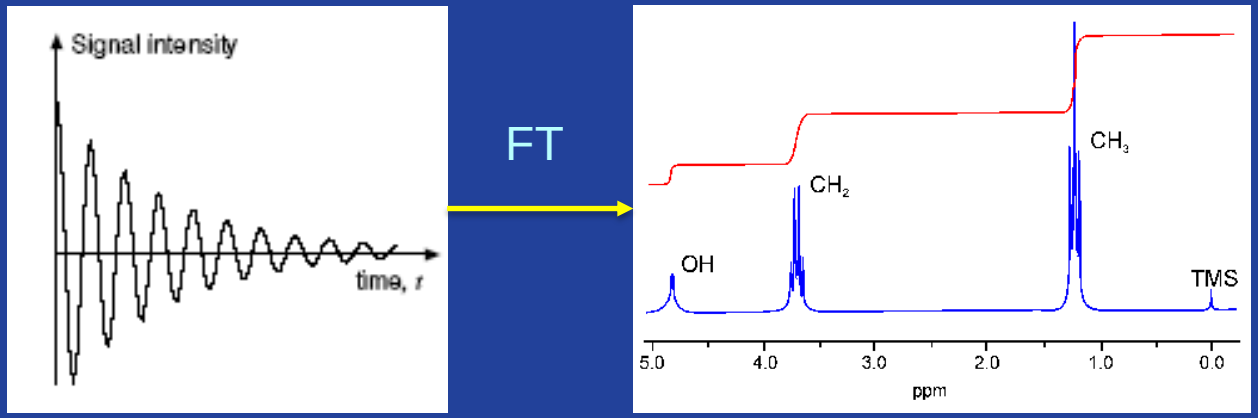
Nuclear magnetic resonance is a *quantum mechanical* property of magnetic nuclei when placed in a strong magnetic field



In a complex molecule, each nucleus experiences a slightly different magnetic field due to local interactions. It is thus possible to detect individual nuclei by measuring the frequency of the energy emitted when a nucleus relaxes from the high energy state to the low energy state

How NMR is measured

Excite all the nuclei in the molecule with a powerful RF pulse, then record the response



Fourier transformation of the response yields the frequency (energy) spectrum.

Very intense magnetic fields are required, produced by superconducting magnets

NMR is a versatile analytic method

Many biomolecular applications of NMR:

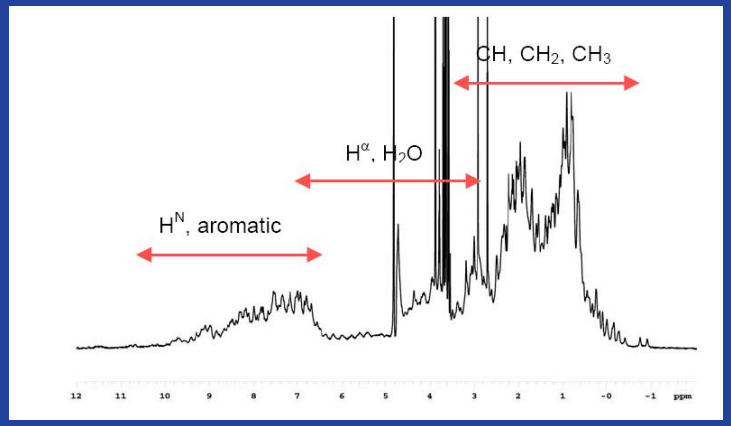
- Structural biology
- Rate processes (dynamics, kinetics)
- Metabolomics
- Drug discovery

This versatility is reflected in the broad range of computer software for NMR

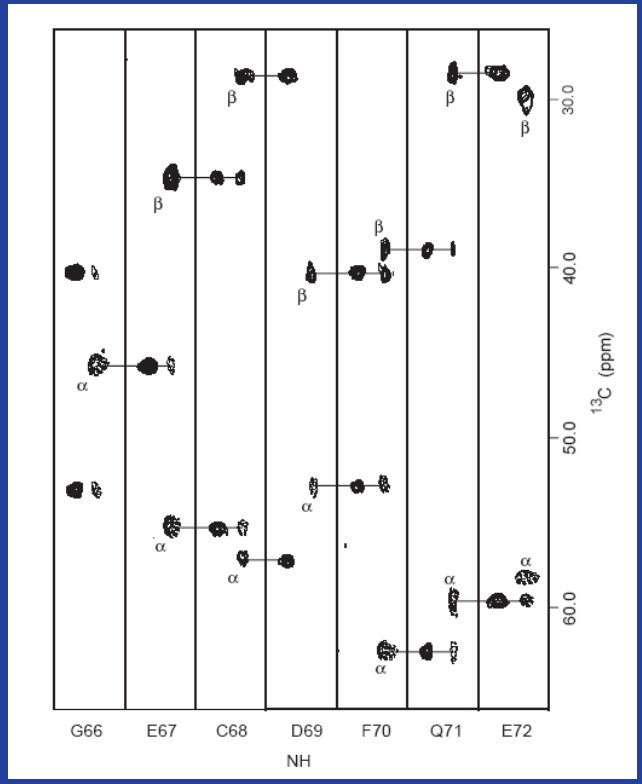
BioMagResBank depositions cite >100 software packages

NMR of proteins: spectral assignment

- Many hundreds to thousands of atoms
- Multidimensional NMR experiments are needed to resolve separate signals for individual atoms



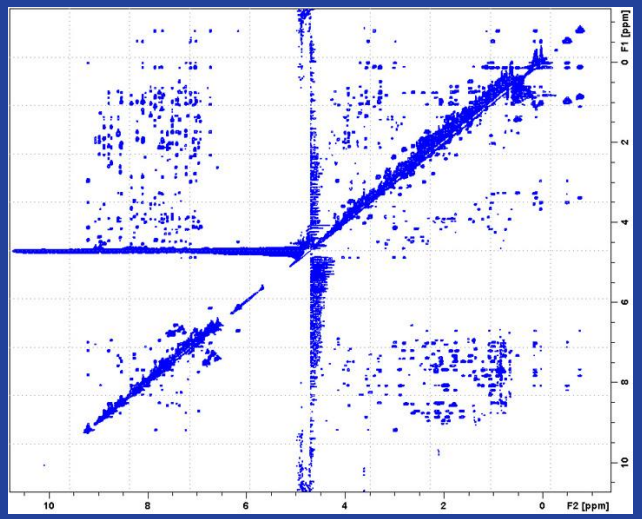
Correlation of the $C\alpha^{i}$ & $C\alpha^{i-1}$ and $C\beta^{i}$ & $C\beta^{i-1}$ sequentially aligns each pair of NHs in the protein's sequence.



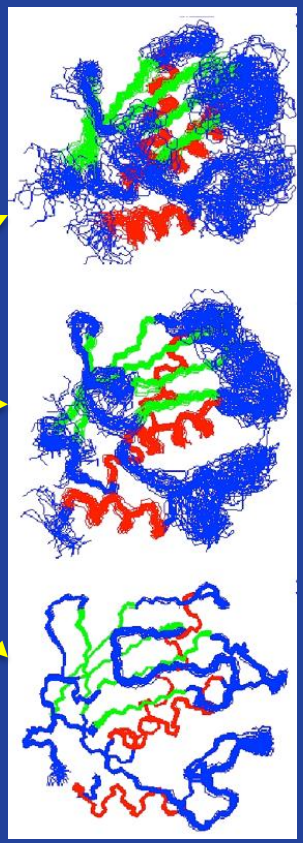
- It is possible to sequentially assign signals to specific amino acids in the protein

NMR of proteins: structure determination

- Short-range distances between atoms can be determined from nuclear Overhauser effect measurements
- Using many of these distances, the protein structure can be determined

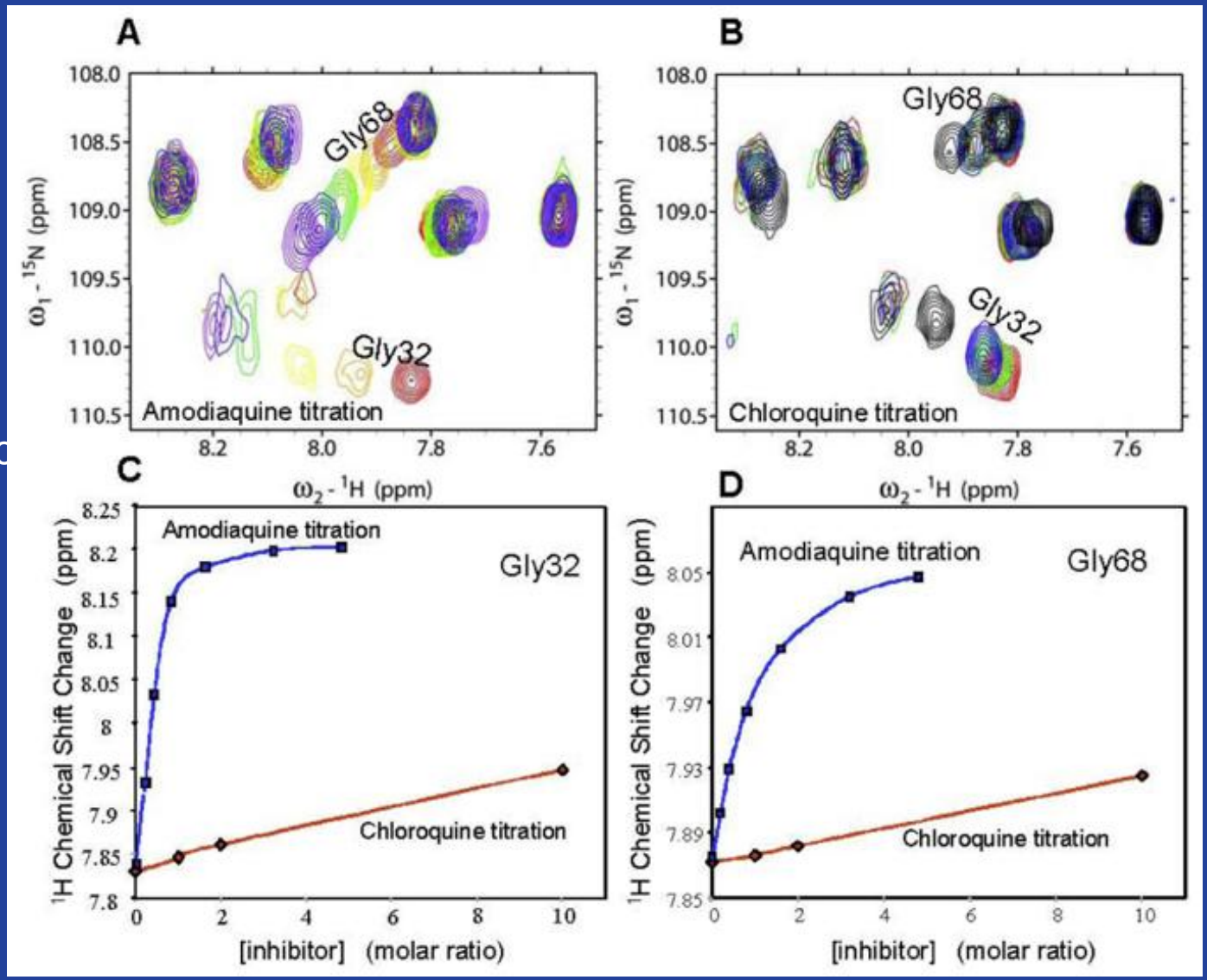


Iterative cycle of refinement as additional peaks are assigned



NMR in drug discovery

NMR rapidly distinguishes specific and non-specific inhibitors



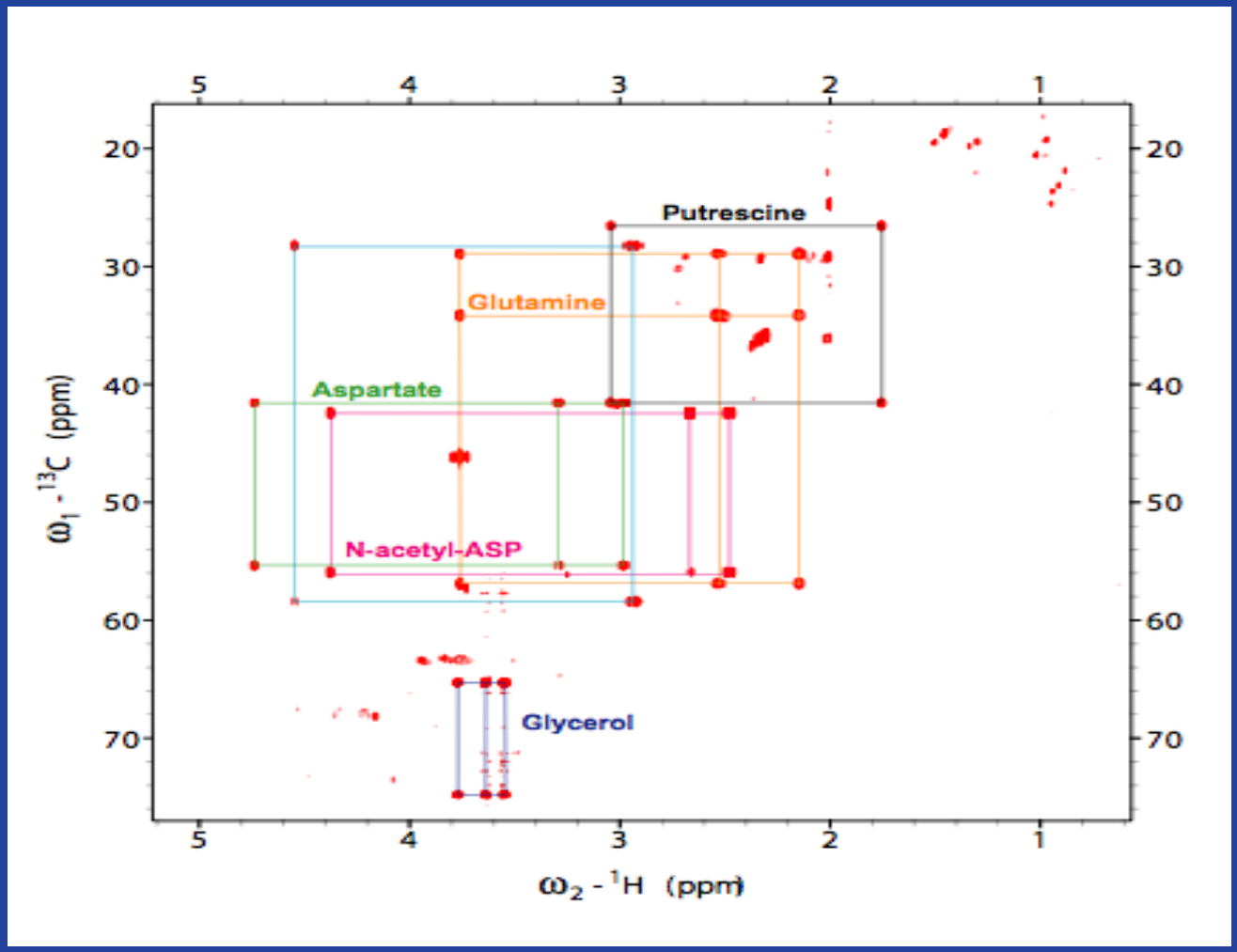
NMRbox

NMR metabolomics

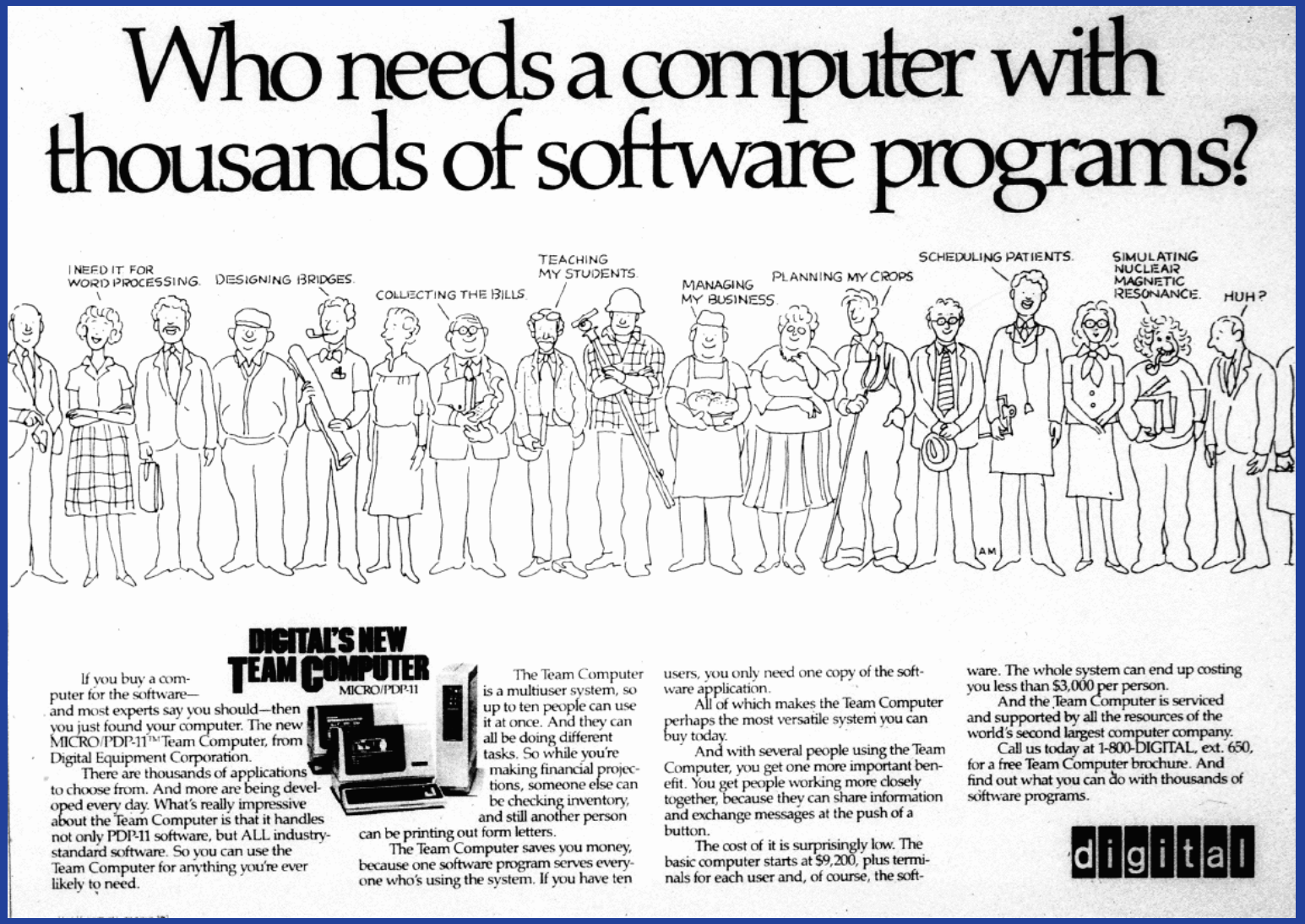
Concentration of all the metabolites in a cell is a direct readout of the cell state

NMR spectrum of E.coli grown on ^{13}C-labelled nutrient

NMR can be used for living cells, organisms (MRI)



A pervasive problem



Who needs a computer with thousands of software programs?

DIGITAL'S NEW TEAM COMPUTER

MICRO/PDP-11

If you buy a computer for the software—and most experts say you should—then you just found your computer. The new MICRO/PDP-11™ Team Computer, from Digital Equipment Corporation.

There are thousands of applications to choose from. And more are being developed every day. What's really impressive about the Team Computer is that it handles not only PDP-11 software, but ALL industry-standard software. So you can use the Team Computer for anything you're ever likely to need.

The Team Computer is a multiuser system, so up to ten people can use it at once. And they can all be doing different tasks. So while you're making financial projections, someone else can be checking inventory, and still another person can be printing out form letters.

The Team Computer saves you money, because one software program serves everyone who's using the system. If you have ten users, you only need one copy of the software application.

All of which makes the Team Computer perhaps the most versatile system you can buy today.

And with several people using the Team Computer, you get one more important benefit. You get people working more closely together, because they can share information and exchange messages at the push of a button.

The cost of it is surprisingly low. The basic computer starts at $9,200, plus terminals for each user and, of course, the software. The whole system can end up costing you less than $3,000 per person.

And the Team Computer is serviced and supported by all the resources of the world's second largest computer company.

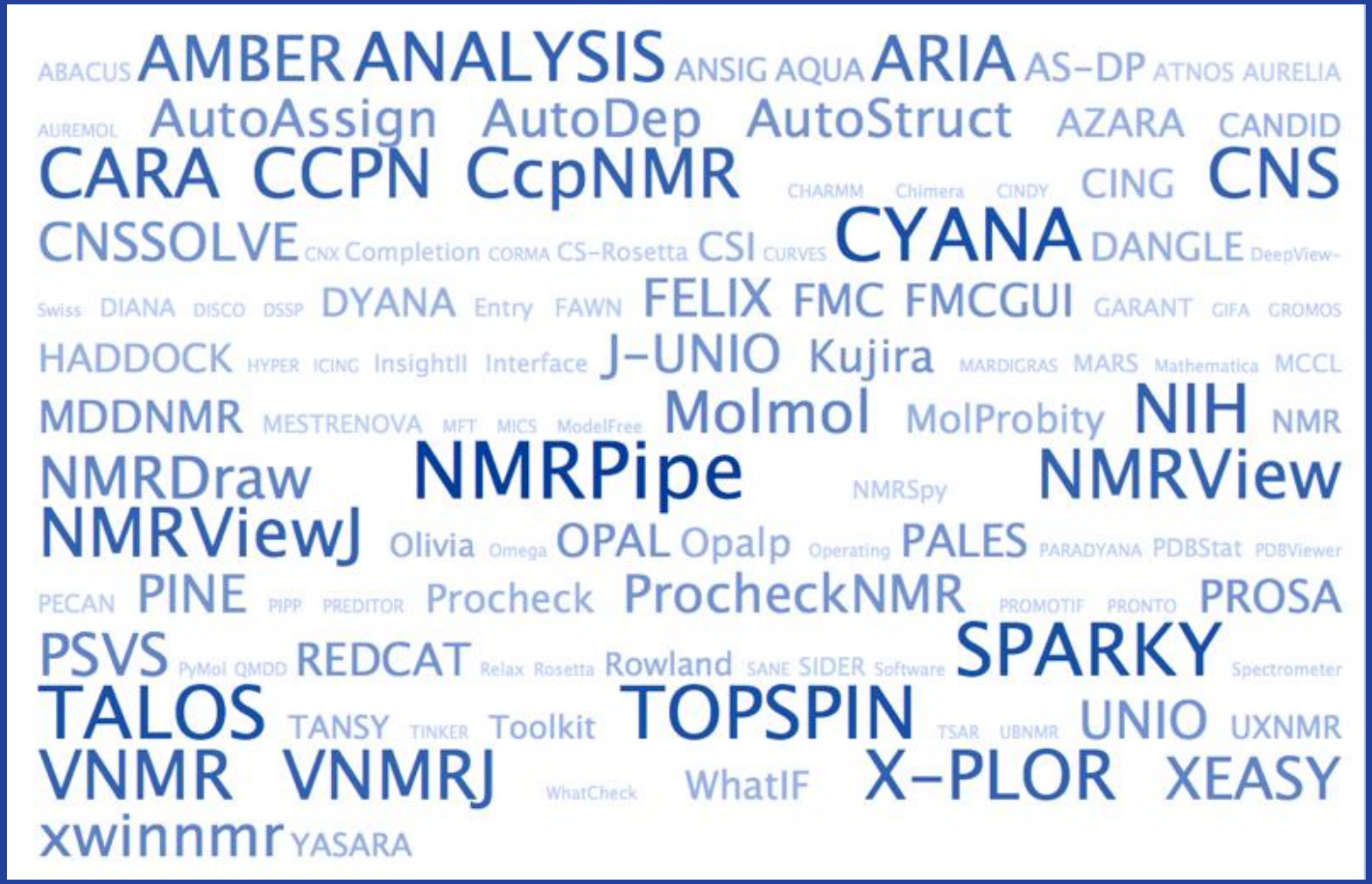
Call us today at 1-800-DIGITAL, ext. 650, for a free Team Computer brochure. And find out what you can do with thousands of software programs.

digital

Advertisement from *Newseek* in 1981

Software abundance

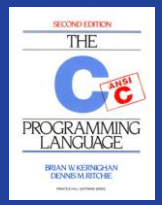
>100 packages cited in BioMagResBank depositions

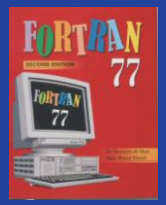


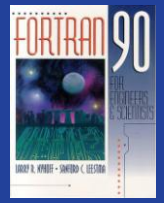
NMRbox

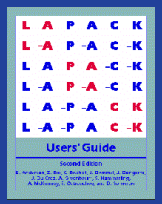
Fragmentation

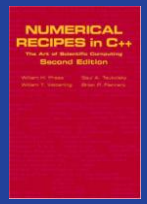
Operating systems, languages, libraries









BLAS

Motivation

- Complex, dynamic software environment
  - Many hundreds of software packages
  - Users: difficult to set up, manage
    (e.g. asynchronous updates, conflicts)
  - Developers: difficult to support (fragmentation)
  - Meta-software ascendant
- Challenges to reproducibility
  - Lack of standards for scripted workflows
  - Lack of annotation for manual steps
  - Poor software persistence

Reproducibility

Seminal paper by Ioannidis, 2005

PLOS | MEDICINE

Essay

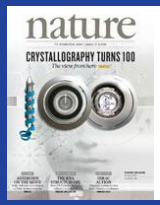
Why Most Published Research Findings Are False

Pharma concerns

MERCK Be well

NOVARTIS

NIH concerns



Policy: NIH plans to enhance reproducibility

Francis S. Collins & Lawrence A. Tabak

27 January 2014

Francis S. Collins and Lawrence A. Tabak discuss initiatives that the US National Institutes of Health is exploring to restore the self-correcting nature of preclinical research.

>70% of published studies on putative drug targets are not reproducible

Obstacles to reproducibility

A computational study is reproducible when it provides the "complete software environment needed to reproduce the figures" (D. Donoho, Stanford)

- Missing primary data
- Missing meta-data
- Missing software (scripts, programs)
- Non-persistence of software
- Manual interventions

Persistence

Why NMR software isn't persistent

Developers graduate

Platforms become obsolete

Grants end

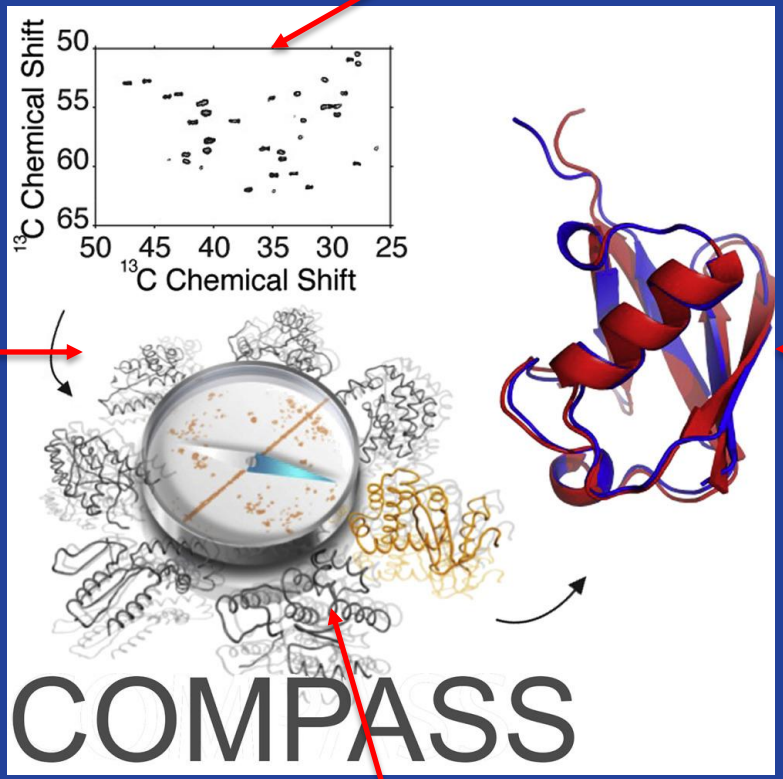
Examples: Sparky, MARDIGRAS, MOLMOL, XEASY, ANTIOPE,…

Meta-software

Experimental Protein Structure Verification by Scoring with a Single, Unassigned NMR Spectrum. Courtney, Rienstra et al. , 2015

Sparky

SHIFTX/X2



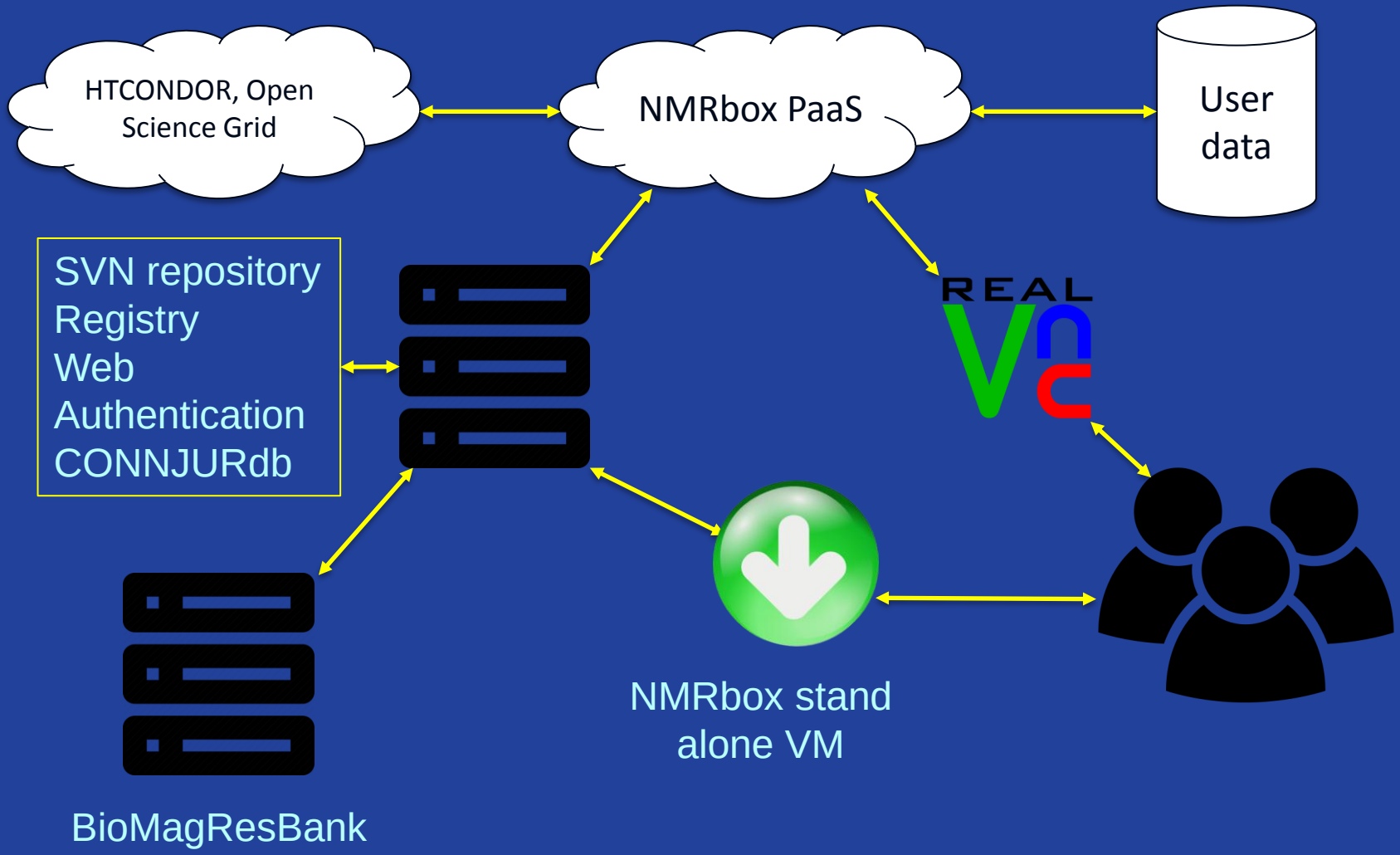
Rosetta

MODELLER

NMRbox approach

- Capture the complex, evolving NMR software environment
  - Delivery via pre-configured, fully provisioned VMs
  - Agnostic: provision with all available software
  - Archive VMs at regular intervals
  - Software registry for discovery
- RDB for workflows (CONNJUR)
  - Regularization
  - Annotation
  - Interoperability
  - Preferred/recommended workflows
- Robust inference
  - Bayesian tools for users
  - API for developers

Functional overview



Registry

NMRbox ENC Workshop Documentation Downloads Registry

Software Registry

GUARDD

guardd
Graphical User-friendly Analysis of Relaxation Dispersion Data

Principal citations:

GUARDD: user-friendly MATLAB software for rigorous analysis of CPMG RD NMR data.
Kleckner IR, Foster MP.
J Biomol NMR. 2012 Jan;52(1):11-22. doi: 10.1007/s10858-011-9589-y. Epub 2011 Dec 11.
PMID: 22160811

Synopsis

GUARDD is a flexible, user-friendly package for analysis of relaxation data built on the power of Matlab. Features include

- Accurate site-specific fits to two-state exchange model at any timescale
- Refinement of best fit for maximum accuracy
- Examine dynamic motion of entire molecule
- Optimization and education with RD Simulator

Keywords

Relaxation analysis, CPMG, Relaxation dispersion

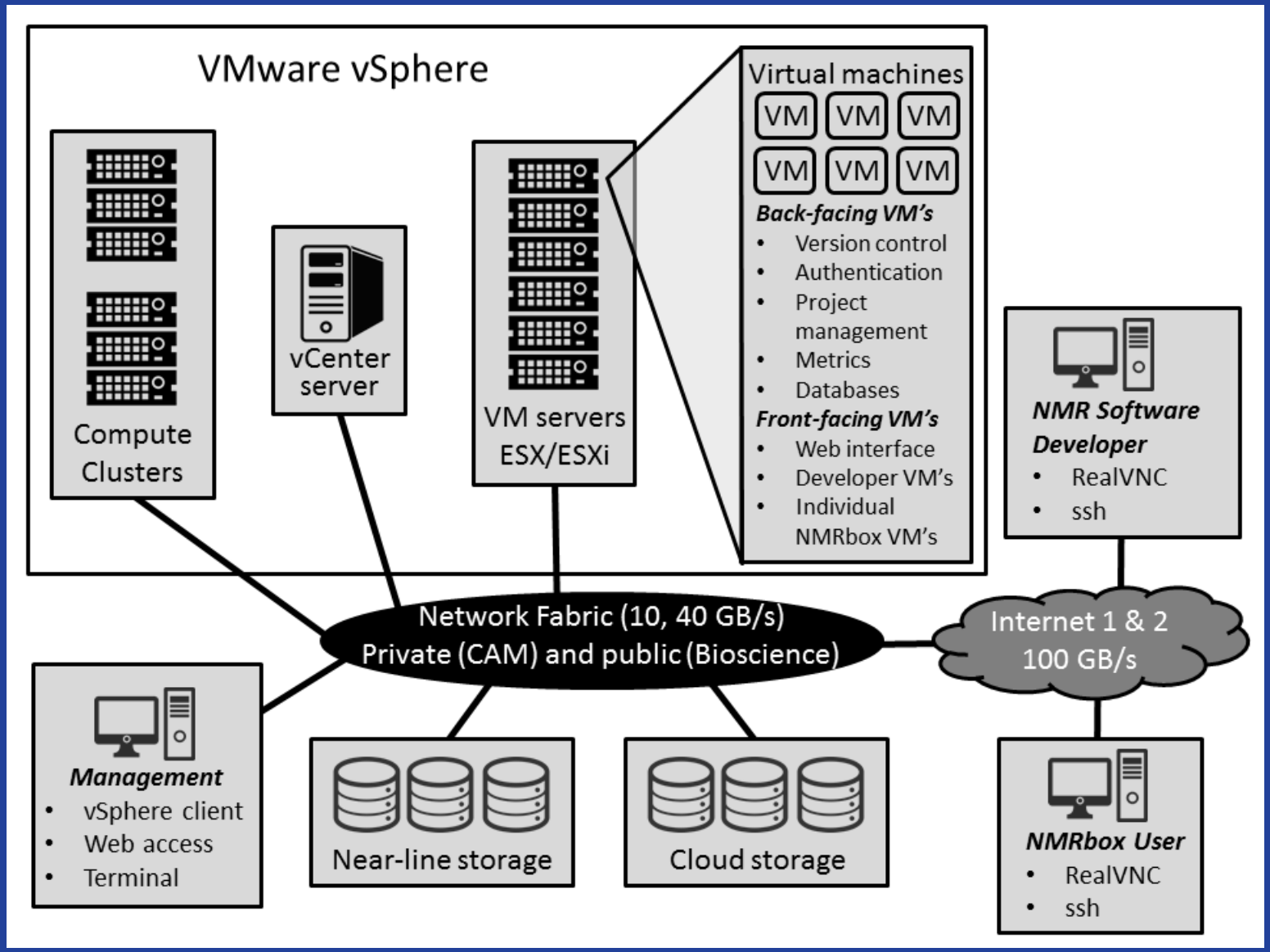
Documentation GUARDD

Related software

DASHA
FAST ModelFree
ModelFree
Relax

BMRB entries citing GUARDD

Infrastructure



Broader Impacts

- NMRbox can serve as a platform for software delivery in other, related domains: X-ray, simulation, thermodynamics
- VMs can serve as a platform for reproducible research in other domains
- Reducing software administration burden allows scientists to spend more time on science
- Reducing hardware requirements allows scientists to spend more $$$$ on science

Unmet challenges

Executable codes with built-in vulnerabilities

- Time “bombs”
- External services
- Hardware dependencies (GPUs, graphic cards)

Making web services persistent and reproducible

- Requires development and acceptance of standards
  - Versioning
  - Mirror
  - Repository (escrow, conditional)

Finding a harmonious solution…

fin