Knowledge discovery using data mined from Nuclear Magnetic Resonance spectral images

William J Brouwer[1], Saurabh Kataria[2], Prasenjit Mitra[2], Karl Mueller[1], C. Lee Giles[2]

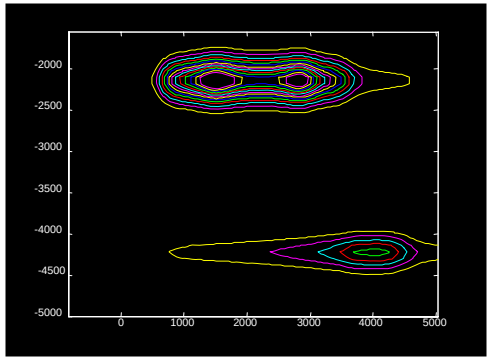
[1]Department of Chemistry, [2]Department of Information Sciences and Technology,

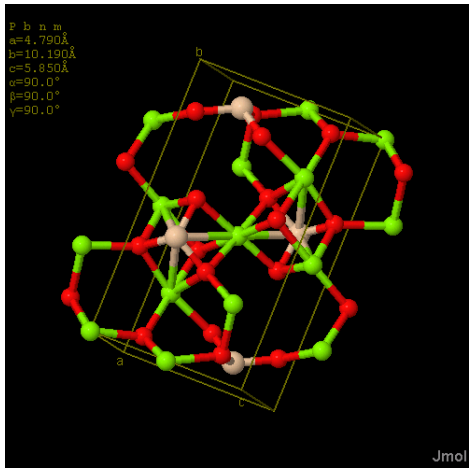
Pennsylvania State University, University Park PA 16802

CHE- 0535656

Outline

- Motivation
  - Structure determination
  - Invoking + building cyberinfrastructure
- Method
  - Solid State *ab initio* calculations
  - Nuclear Magnetic Resonance (NMR)
  - Support Vector Machines (SVM) + NMR
- Results
  - High Resolution experiment + Ensemble SVM
- Future Work
  - Sequestration
  - Surface science
- Conclusions





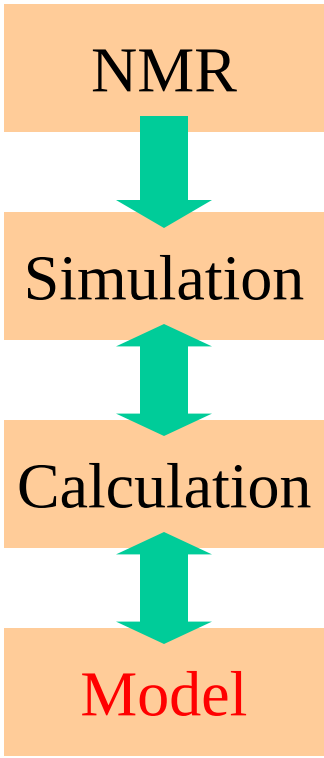
Motivation

- Solid State NMR is powerful local probe of atomic structure
  - Local geometry; bond angles, lengths
  - Local chemical identity (eg., steric differences directly influence local bonds etc)

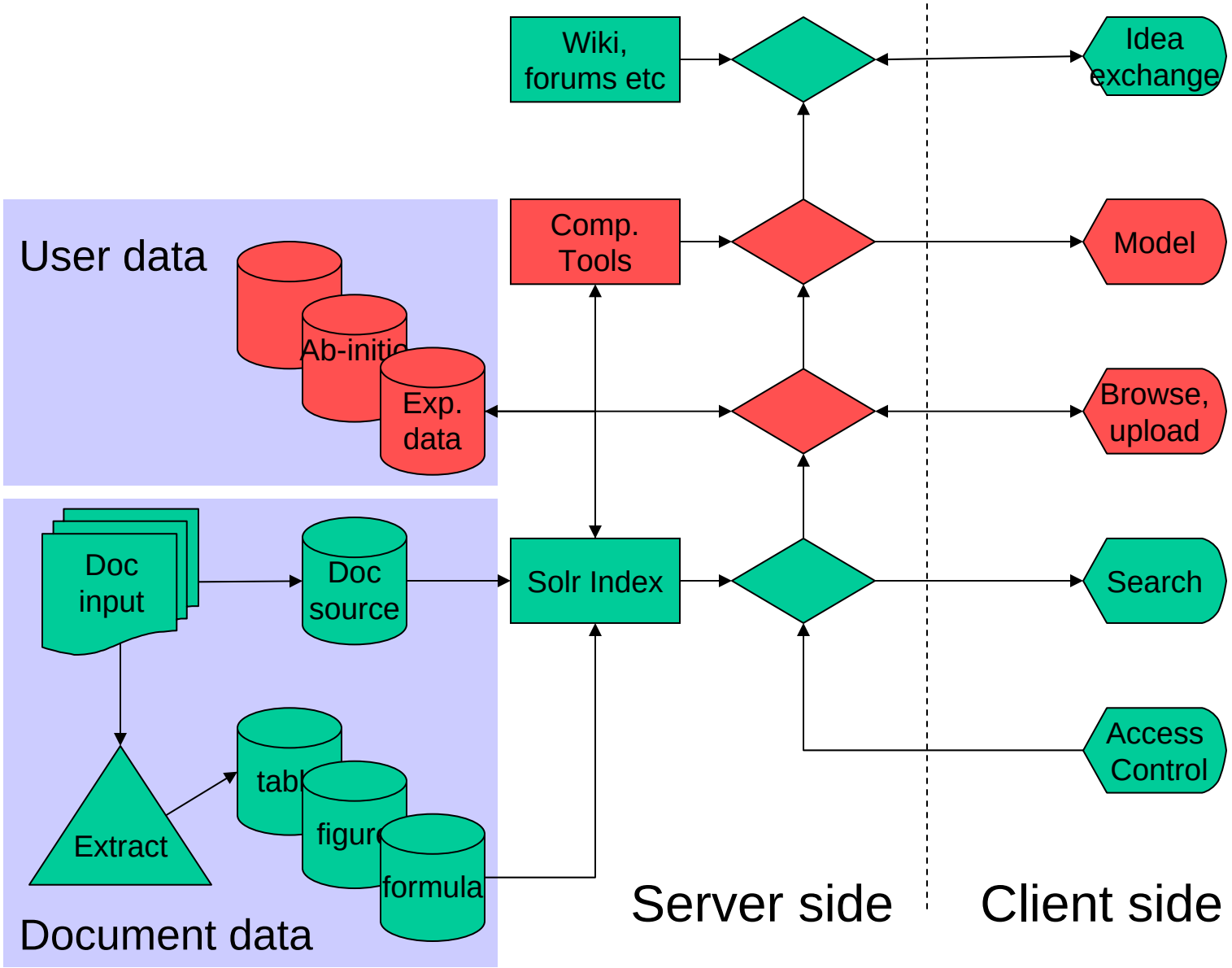
- NMR Lineshapes often complicated by broadening mechanisms, *interpretation requires intensive work including simulation, ab initio and/or empirical calculations*

- Machine Learning (ML) promises to reduce the burden of interpretation by *removing* intermediate steps

- Requires voluminous *data*, such as that provided by cyberinfrastruture, eg., $Chem_XSeer$ -> http://chemxseer.ist.psu.edu/

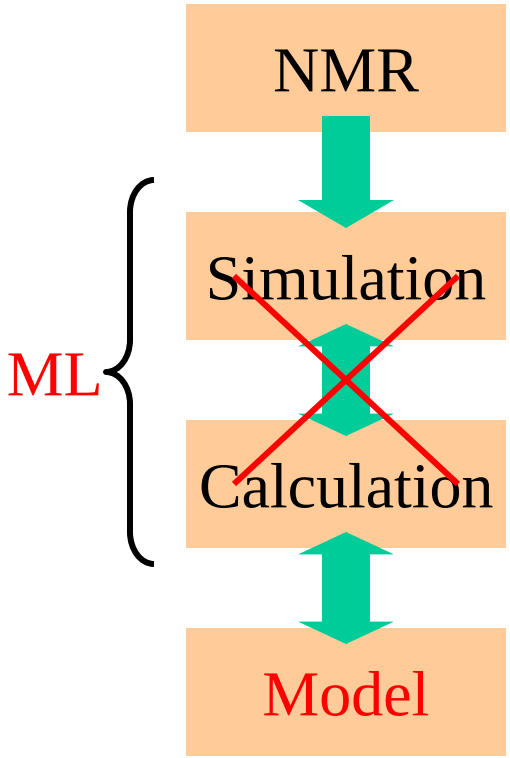


The ChemXSeer Collaboratory



Method

- Collaboratories bridge divides imposed by resources, geography to create a distributed research environment
- Chem$_X$Seer provides access to digital libraries and allows end-users to search using unique features such as tables, figures as well as text from documents
- *Using data from Chem$_X$Seer, an end-user may employ machine learning to determine structural models for NMR spectra of novel materials*
- The focus of this work is on using data from *ab initio* calculations and simulations of NMR spectra, to train support vector machines, to be used for structure determination



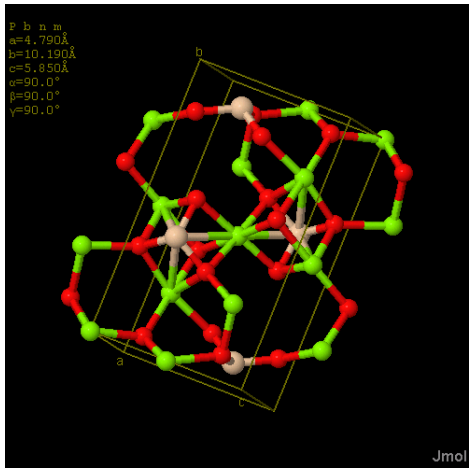
Solid State *Ab Initio*

- Solve Schrodinger (many body) equation with approximations eg., Born-Oppenheimer, Kohn-Sham DFT, to find *electronic wave functions* ϕ, given input structure/unit cell*:*

$$\left\{-\frac{\hbar^2}{2m}\nabla^2 + u_{ext}(\mathbf{r}) + \int \frac{\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}' + \mu_{XC}[\rho]\right\}\varphi_i(\mathbf{r}) = \varepsilon_i\varphi_i(\mathbf{r})$$

- Use plane wave (PW) expansions + pseudopotentials for core regions, accounting for full crystal periodicity.
- Can also relax structure ie., find optimal atomic positions by minimizing forces between atoms
- Goal of *ab initio* methods is to calculate physical properties we may measure by some experimental method

Forsterite Unit Cell



Example: Forsterite

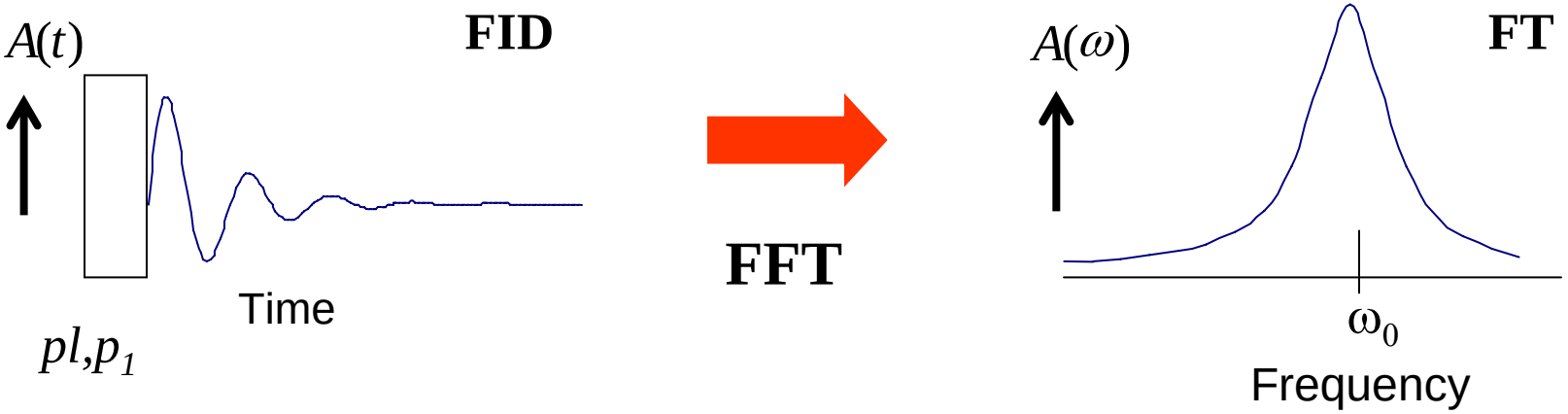
- Construct pseudopotentials for Si,O,Mg eg., for Mg pseudize core of $1s^2 2s^2 2p^6$, valence orbitals $3s^2 3p^0 3d^0$ treated with Projector Augmented Waves (PAW)
- Using refined (and/or relaxed) structure, perform Self-Consistent Field calculation for electronic structure, using DFT
- Resultant wave functions ψ take into account full system periodicity, *may be used to calculate expectation values of electric field gradient tensor* **V** *measurable in solid state NMR via the quadrupole coupling constant* C_q *and asymmetry parameter* η*:*

$$V_{zz} = \langle \psi | \mathbf{V}_{zz} | \psi \rangle;$$
$$V_{yy} = \langle \psi | \mathbf{V}_{yy} | \psi \rangle;$$
$$V_{xx} = \langle \psi | \mathbf{V}_{xx} | \psi \rangle$$

$$C_q = \frac{e^2 V_{zz} Q}{\hbar}; \eta = \frac{V_{yy} - V_{xx}}{V_{zz}}$$

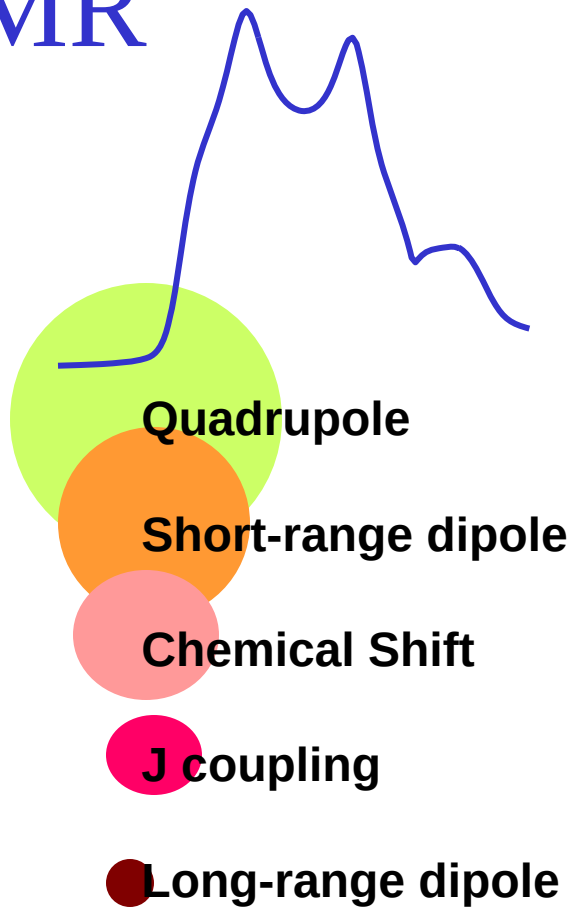
Solid State NMR (SS-NMR)

- Nuclei with non-zero spin *I* couple with static magnetic field **B** to produce $2I+1$ Zeeman energy levels; nuclear spin precess at Larmor frequency ω_0
- Samples embedded in static **B** field are pulsed with RF energy at ~ Larmor freq, with pulse duration p_1 and power *pl*
- Time response is Free Induction Decay, Fourier transformed via FFT; peak positions (shifts away from Larmor freq.) are functions of *interactions* and thus local structure…



Interactions of NMR

- Vast majority of nuclei have spin > ½ and thus quadrupole moment *Q* which couples with a surrounding electric field gradient, largest interaction
- In polycrystalline solids, this *quadrupole interaction* broadens NMR lines :
  - helpful b/c distinctive lineshapes give direct insight into local bonding arrangement
  - detrimental b/c promotes overlap between lines from distinct chemical environments & need to simulate
- Mitigate with Magic Angle Spinning (MAS), mechanical technique which on time average removes 1st order quadrupole broadening, but 2nd order effects remain…



Multiple Quantum MAS (MQMAS)

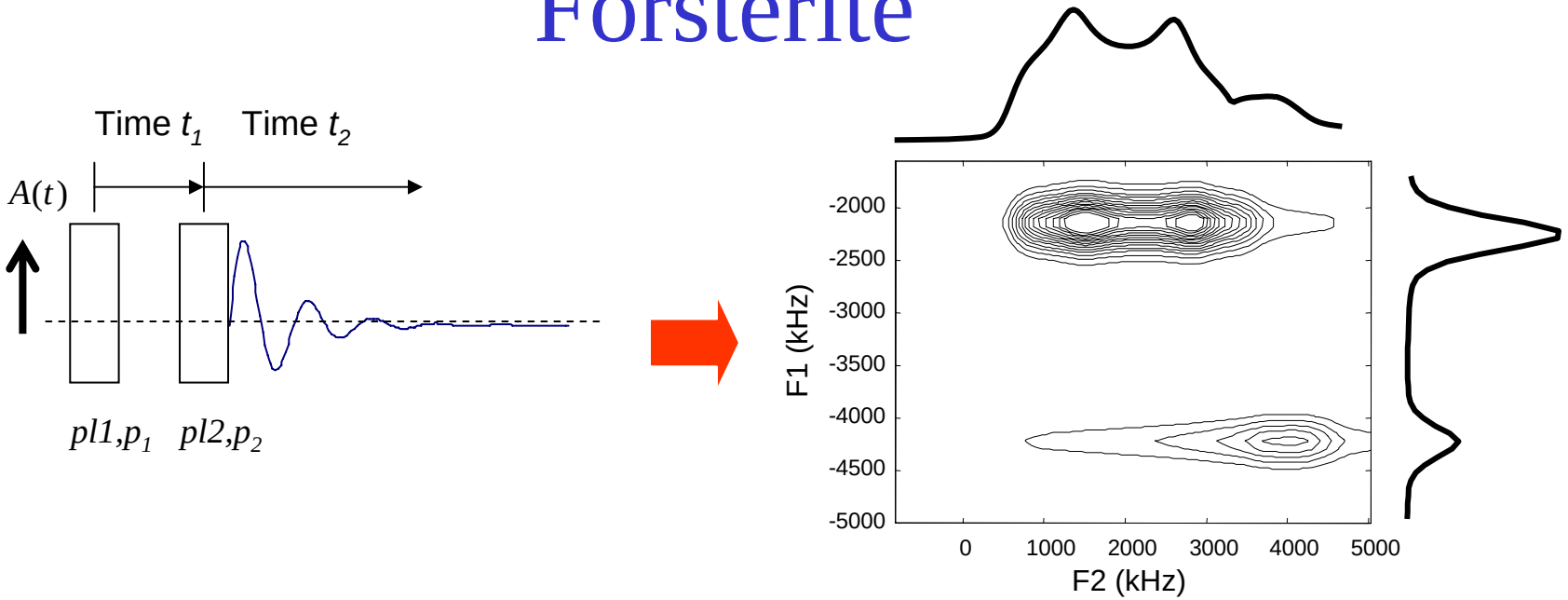
- Bary centers for quadrupole nuclei (I>1/2) using MAS have an isotropic chemical and quadrupole shift, as well as an anisotropic term, which introduces broadening in polycrystalline materials, function of crystallite orientation α,β :

$$\omega_{r,c} = (r-c)\omega_0\delta_{cs}^{iso} - \frac{r-c}{\omega_0}\left[\frac{C_q}{2I(2I+1)}\right]^2\left\{A^{(0)}(I,r,c)\left(\frac{\eta^2+3}{10}\right) + A^{(4)}(I,r,c)\,f(\eta,\alpha,\beta)\right\}$$

Chemical shift — 2nd order quadrupole shift — Anisotropic term

- Can only detect coherences with change in magnetic quantum number (energy transitions) with r-c = +/- 1 (single quantum coherence)
- MQMAS -> acquire data in experiment as function of *two* independent time intervals, multiple quantum coherences detected indirectly (evolve btwn pulses)

Example: ^{25}Mg MQMAS of Forsterite

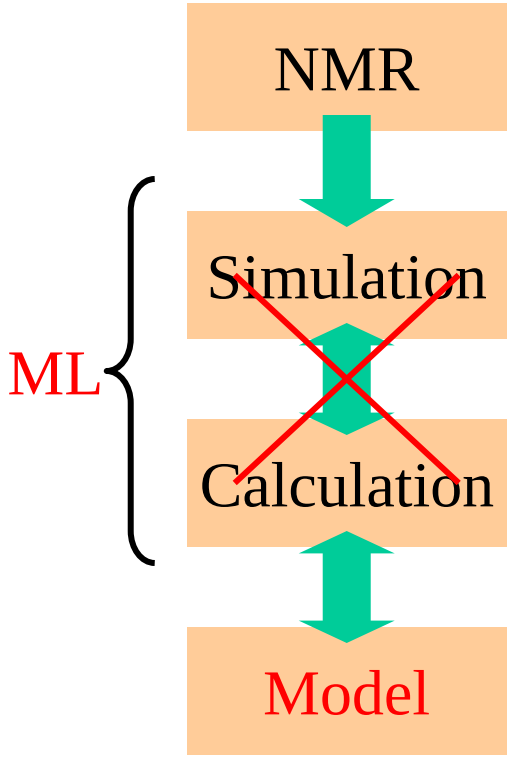


- FFT in two dimensions, indirect dim is *isotropic*, by virtue of experimental details and/or data processing

- Two inequivalent Mg sites revealed as distinct NMR peaks, having specific values for asymmetry parameter η and quadrupole coupling constant C_q

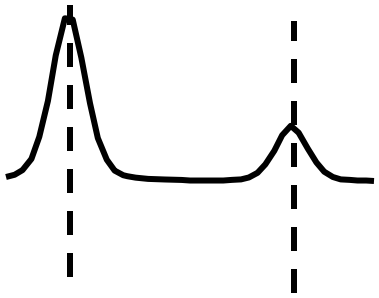
Machine Learning with SS-NMR data

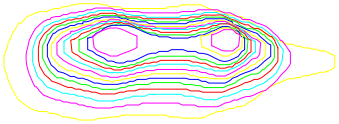
- Comparison of parameters extracted via simulation of spectrum with *ab initio* results allows for unequivocal assignment
- Ideally would like to eliminate/reduce ab initio + simulation in order to do interpretation on SS-NMR spectra
- 2D data may exist as either image (eg., from article) or processed binary data from experiment
- In the former case, pixel data representing data easily maps to intensity *IF* contour levels are defined (eg., in figure caption)
  - contours don't overlap, use methods developed in ChemXSeer with heuristics + CCL to recreate data
  - Couple with document text and/or ab initio to provide corresponding structure for input spectra…



Features for Machine Learning

- Need position + intensity invariant features for ML



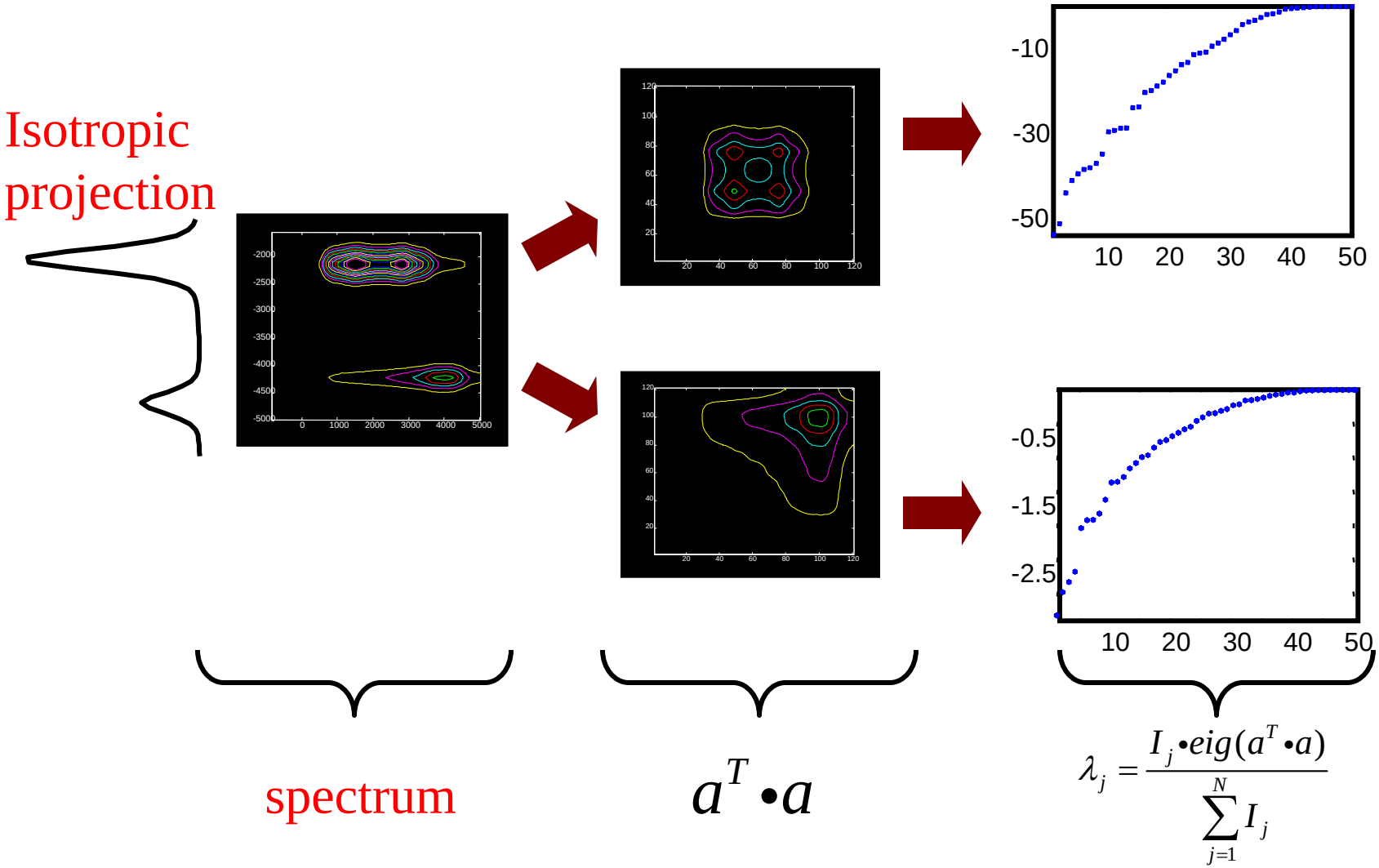


$$a = \begin{matrix} 3.12 & 3.15 & \cdots & 2.97 \\ 3.15 & 3.17 & & 2.63 \\ \vdots & & & \vdots \\ 2.33 & 1.91 & \cdots & 4.12 \end{matrix}$$

$$\lambda_j = \frac{I_j \bullet eig(a^T \bullet a)}{\sum_{j=1}^{N} I_j}$$

- Use isotropic dimension in MQMAS to locate peaks for distinct chemical sites
- Intensity data may be large and many in number eg., 200x100 frequency points = 20k features
- Use eigenvalues of Hessian scaled by relative intensities *I* to reduce dimensions/no. of features

Example: Features from ^{25}Mg MQMAS, Forsterite



Support Vector Machines

- Machine Learning for overlapping decision regions traditionally used back propagation ANN.
- Last decade SVM has proven popular owing to computational tractability; given set *G* of input $\{x_i\}$ output $\{a_i\}$ data, non-linear functions ψ_i map input data to higher dim feature space.
- SVM regression function has weights w_i and constants *b*, adjusted via *regularized risk minimization:*

$$f = g(x) = w_i \psi_i(x) + b$$

- *Using features extracted from NMR spectra* $\{a_i\}$ *in conjunction with structural details gleaned data* $\{x_i\}$ *one may perform structure prediction…*

Ensemble SVM

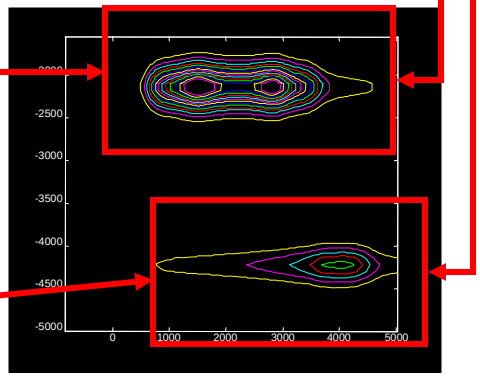
- Assign a single SVM to each atomic position component eg., for Forsterite unit cell, 6atoms x 3 (xyz) = 18 SVM's
- There are two distinct Mg sites in forsterite MQMAS spectra -> use ensemble of *two* SVM grids, weighted by training accuracy
- *Input:* (each grid element) 50 eigenvalues
- *Output*: (each grid element) one atomic position component

Atomic Positions

Si_x	Si_y	Si_z
$1Mg_x$	$1Mg_y$	$1Mg_z$
$2Mg_x$	$2Mg_y$	$2Mg_z$
$1O_x$	$1O_y$	$1O_z$
$2O_x$	$2O_y$	$2O_z$
$3O_x$	$3O_y$	$3O_z$

$$\begin{matrix} & SVM_{2,1} & \cdots & SVM_{2,3} \\ SVM_{1,1} & \cdots & SVM_{1,3} & \vdots \\ \vdots & SVM_{2,16} & \cdots & SVM_{2,18} \\ SVM_{1,16} & \cdots & SVM_{1,18} & \end{matrix}$$

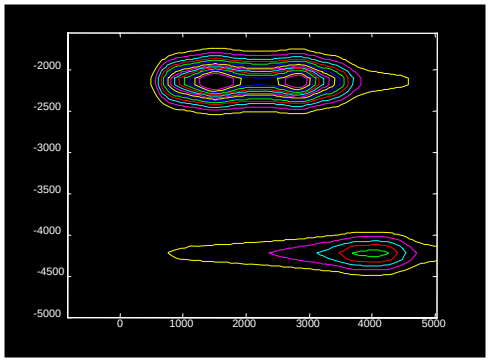
Ensemble SVM

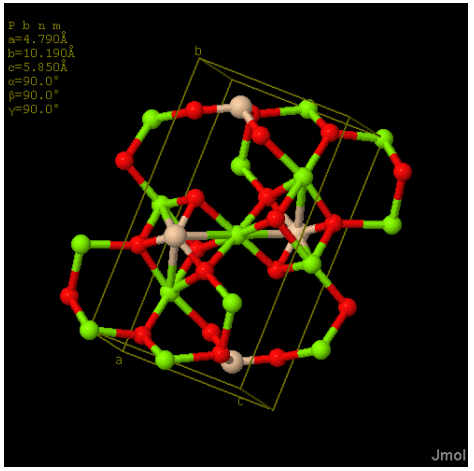


^{25}Mg MQMAS

Results

- 50 sets of input structures produced with random displacements of < 5%
- *Ab initio* performed in ESPRESSO, custom software used to simulate MQMAS spectra using calculated C_q and η
- SVMlight is used for ensemble creation/training, using 2*50*50 features, matched with atomic coordinate components
- *Ensemble processing MQMAS spectra produced from trial structures reproduce structure to within 10%*
- Accuracy should increase further with larger training set, more accurate ab initio (these abbreviated runs ~ 20mins each on dual Xeon)





Current Work

- Procedure as outlined works for when some insight exists as to structure eg., useful for understanding structural phase changes
- A large amount of work is devoted to understanding uptake of heavy elements in minerals
- Ideally would like to extend this method to predicting substitution sites and percentage uptakes of Cs,Sr in (for example) zeolites, clays etc
- Application to disordered materials eg., glasses, solid solutions
- Surface science eg,. studies of liquid/solid interface etc

Conclusions

- Solid State NMR is powerful local probe of atomic structure, impeded by need for simulation, ab initio and/or empirical calculations
- Collaboratories provide wealth of information in the form of experimental, calculated and document data such as figures, tables etc
- Machine Learning techniques eg., SVM may be trained on said data using features described herein, to give direct insight into local atomic structure.
- Methods outlined have promising applications besides structural morphology eg., sequestration

Acknowledgements

- Microsoft
- National Science Foundation

References

•T. Joachims, **Making large-Scale SVM Learning Practical.** Advances in Kernel Methods - Support Vector Learning, B. Schölkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.

•**ESPRESSO** http://www.quantum-espresso.org/

•Brouwer, W. J; Davis, M. C.; Mueller, K. T., **Optimized Multiple Quantum MAS Lineshape Simulations in Solid State NMR** *Computer Physics Communications* (Submitted).