Predicting Phase of Simulated Molecules Using Radial Structural Functions

Heejung Chung (hchung98@stanford.edu), advised by Rodrigo Freitas*

*Reed Materials Computation and Theory Group

Background

Radial Structural Function (RSFs)

- For ith molecule,

$$G^{(i)}(r,\sigma) = \sum_{j=1}^{n} e^{-(d_{ij}-r)^2/2\sigma^2} \quad (1, 4)$$

$G^{(i)}(r) \approx$ # neighbors that are r Angstroms away

- E.g. if red molecule in Fig 1 is ith, then
$G^{(i)}(\mathbf{r}_1) \approx 4$
$G^{(i)}(\mathbf{r}_2) \approx 5$
$G^{(i)}(\mathbf{r}_3) \approx 3$
$G^{(i)}(\mathbf{r}_4) \approx 4$

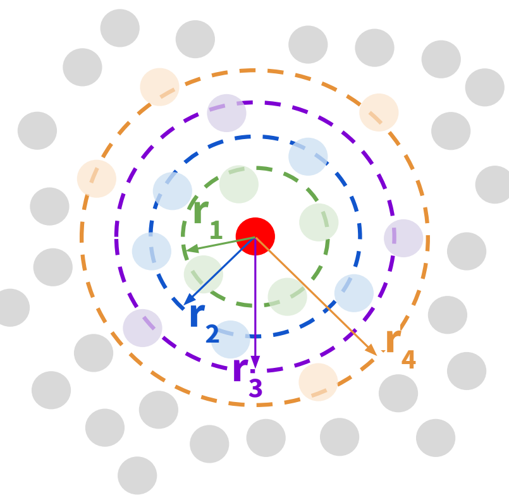


Fig 1: diagram to illustrate approximating RSFs (adapted from figure by Rodrigo Freitas)

- Can be obtained only from simulation

Radial Distribution Function (RDF)

- Scaled mean of RSFs (averaged over all molecules in a sample)
- Obtained experimentally, but produced here from simulation

Predicting

Predicting Phase

Predicting phase

- *Motivation:* No current consensus on how to featurize material structure, but RSFs are promising [2]
- *Supervised learning problem:* trained SVM on RSFs to predict phase (simple structural property) of simulated H_2O molecules

Generating RSFs

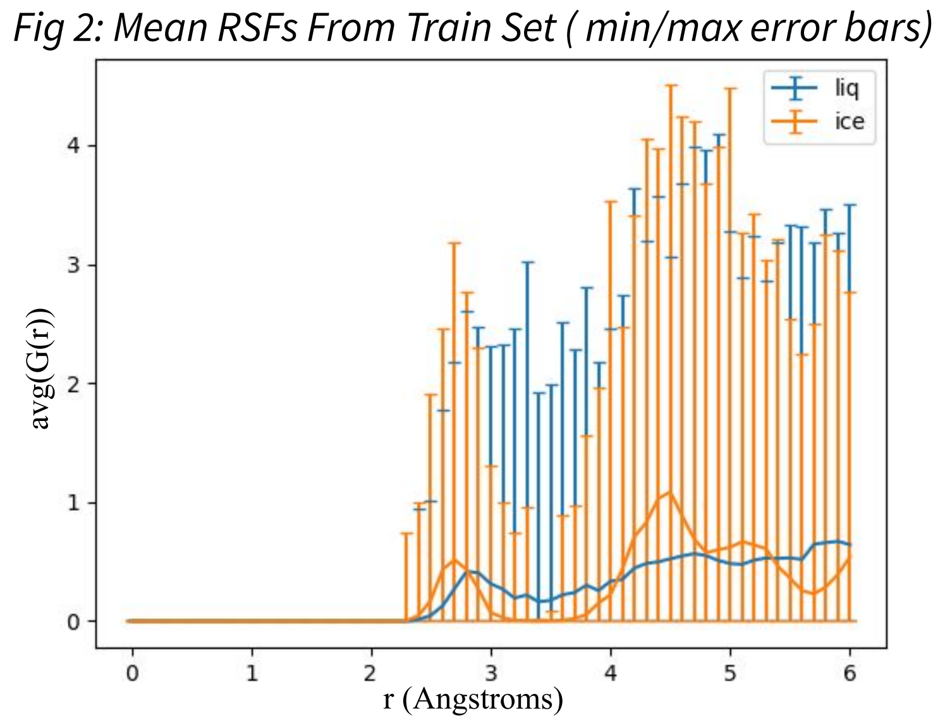
Generating RSFs from RDF

- *Motivation:* Want to use knowledge of RSFs to augment experimental data (RDF) which may be expensive and difficult to obtain
- *Unsupervised learning problem:* Used RDF (scaled mean of RSFs) to generate pseudo RSFs, by sampling from different probability distributions

Data and Features

- Ran two simulations[3] at 250K: one of ice, one of water
  - Computed RSFs of molecules at two timestamps (t1, t2) from each simulation
  - *Train set:* {ice RSFs from t1} ∪ {water RSFs from t1}
  - *Dev set:* {ice RSFs from t2} ∪ {water RSFs from t2}
- *Examples:* correspond to simulated molecules
- *Features:* $x^{(i)}=\{G^{(i)}(0Å), G^{(i)}(0.1Å),...G^{(i)}(5.95Å), G^{(i)}(6Å)\}$
- *Labels:* $y = +1$ if molecule came from water simulation, $y = -1$ if from ice simulation

Fig 2: Mean RSFs From Train Set (min/max error bars)



Models

Predicting phase

- Unregularized Linear SVM (baseline)

$$min_{w,b}||w||^2 \quad \text{subject to} \quad y^{(i)}(w^T x^{(i)} + b) \geq 1,\ i = 1, ..., n$$

- Regularized Linear SVM (optimal regularization term $C=10$, see Fig 4)

$$min_{\gamma,w,b}\frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i \quad \text{subject to} \quad \begin{array}{l} y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i,\ i = 1, ..., n \\ \xi_i \geq 0,\ i = 1, ..., n \end{array}$$

Generating RSFs from RDF

- Sampled from gaussian, assumed...
$G^{(i)}(r_j) \sim \mathcal{N}(\mu_j, \sigma)$
  - Produced *Pseudo Train Set 1* by sampling 768 examples from gaussian parameterized by ice RDF and 768 examples from water RDF
- Sampled from exponential, assumed...
$G^{(i)}(r_j) \sim \mathrm{Exp}(\lambda = 1/\mu_j)$
  - Produced *Pseudo Train Set 2* by sampling 768 examples from exponential parameterized by ice RDF and 768 examples from water RDF
- Assumptions and notes for both
  - Assumed RSFs from different r's are independent
  - Got μ_j for each r_j from RDF (essentially scaled MLE)
  - Trained SVMs on produced *Pseudo Train Sets* and validated using dev set

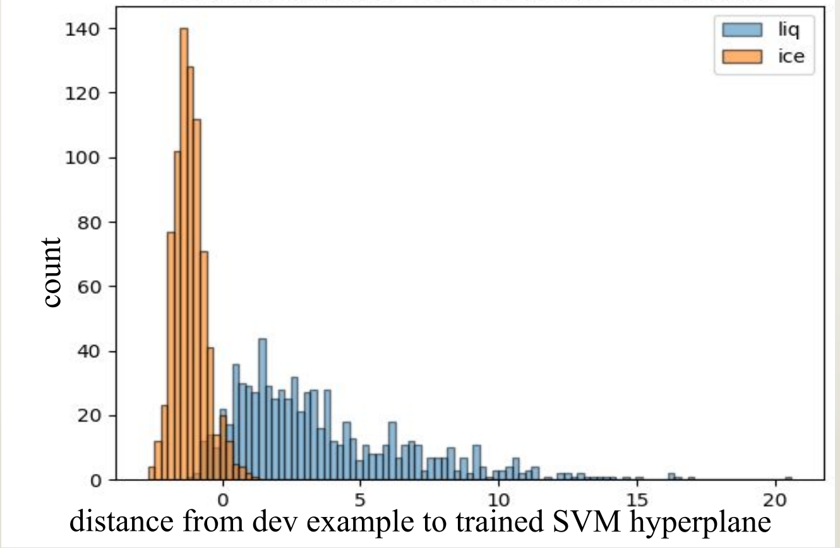
Results

Results: Overview

- For both train and dev sets, n = 1536
  - i.e. each dataset included 768 water, 768 ice molecules

Model	Train Error	Dev Error
unregularized SVM trained on **real** RSFs	5.5%	5.7%
SVM trained on **real** RSFs, C=10	5.0%	5.7%
SVM trained on **pseudo** RSFs (gaussian)	0%	17%
SVM trained on **pseudo** RSFs (exp)	0%	11%

Fig 3: Histogram of Decision Function of Dev Set



Predicting phase

Fig 4: Validation Curve

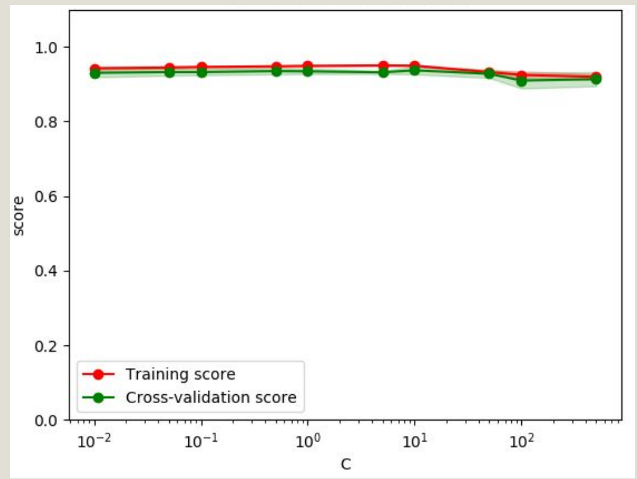
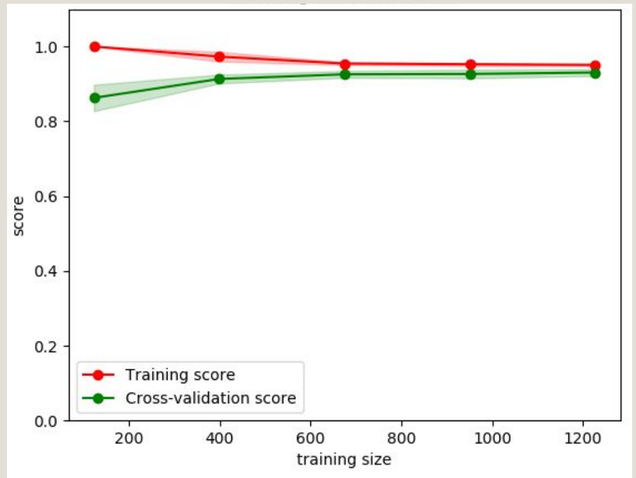


Fig 5: Learning Curve (C=10)



Generating RSFs from RDF

- *Histograms of decision function for different SVMs:* gaussian pseudo RSFs plot (Fig 6) differs more than exp pseudo RSFs (Fig 7) from real RSFs (Fig 3)
- Suggests that exp pseudo RSFs SVM hyperplane is closer to real RSFs SVM hyperplane, i.e. exp pseudo RSFs are distributed more like real RSFs than gaussian pseudo RSFs

Fig 6: Histogram of Decision Function of Dev Set (from SVM trained on gaussian pseudo RSFs)

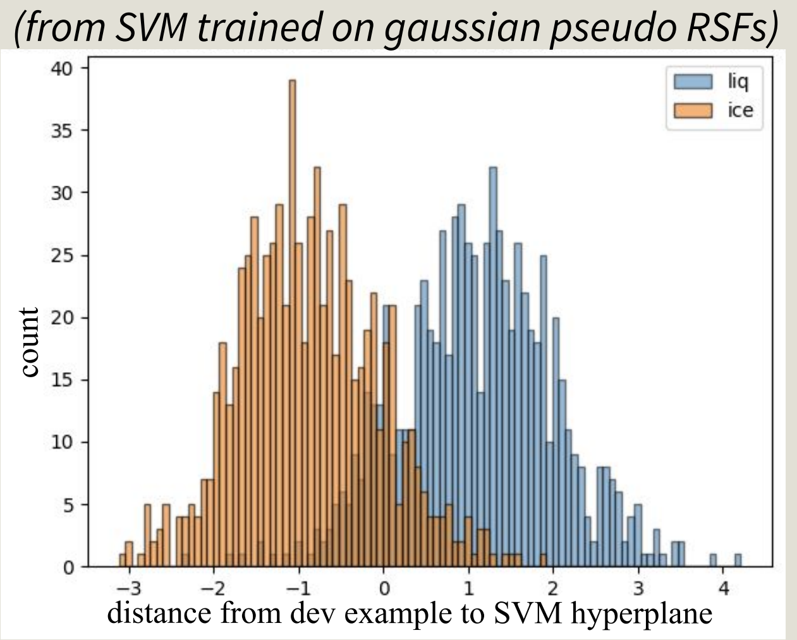
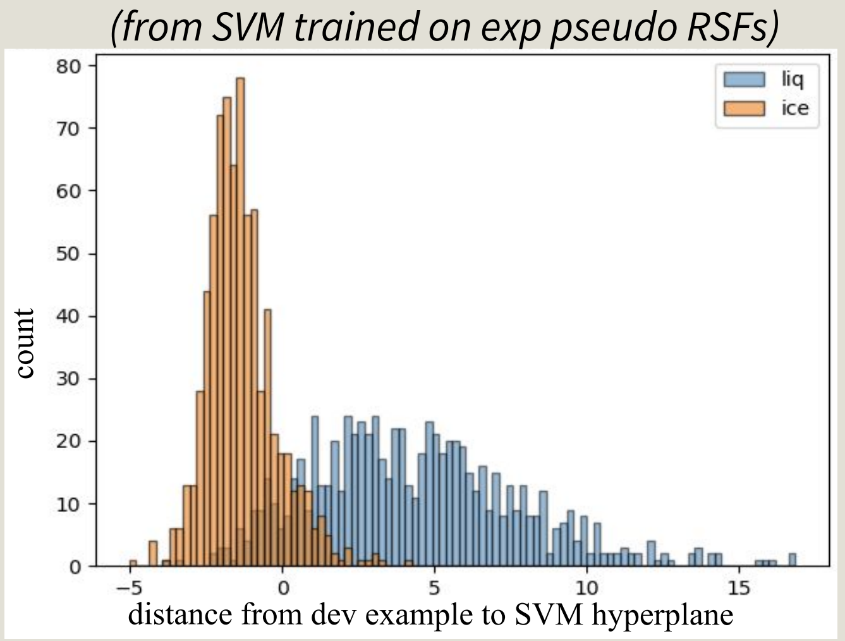


Fig 7: Histogram of Decision Function of Dev Set (from SVM trained on exp pseudo RSFs)



Discussion

- RSFs are good for predicting phase of H_2O, which is promising if we want to use RSFs as featurizations of material structure
- Each RSF $G(r)$ seems to be exponentially distributed, which is unexpected
- Can augment simulated "experimental data" for single-phase substances by sampling RSFs from pdfs parameterized by RDFs

Future

- *Refine pseudo RSF generation*: try sampling RSFs from multivariate gaussian or dirichlet parameterized by RDFs to take correlation between RSFs into account
- *Generate RSFs for mixed-phase:* now that we know how to augment RDFs with pseudo RSFs for pure liquid and pure ice, find model that can generate RSFs from RDF of sample that is mixed (eg 70% water, 30% ice) and predict what portion of sample is composed of water
- *Look at RSFs for more complex structural properties:* for example, analyze RSFs from materials with defects like dislocations and vacancies vs RSFs from material without defects

Citations

1. J. Behler, and M. Parrinello, PRL 98, 146401 (2007)
2. Molinero, Valeria, and Emily B. Moore. "Water modeled as an intermediate element between carbon and silicon." The Journal of Physical Chemistry B 113.13 (2008): 4008-4016
3. S. Plimpton, Fast Parallel Algorithms for Short-Range Molecular Dynamics, J Comp Phys, 117, 1-19 (1995). http://lammps.sandia.gov
4. S. Schoenholz, et al., Nature Physics 12.5 (2016)