CS 6824:
AI-powered Molecular Modeling

Website: https://people.cs.vt.edu/dbhattacharya/courses/cs6824/

Piazza: https://piazza.com/vt/spring2022/cs6824/home

Canvas: https://canvas.vt.edu/courses/145337
Are you in the right place?

- This is CS 6824: CRN 20577
  - Modality is "Face-to-Face Instruction"
Today

- What is AI-powered Molecular Modeling, the field, about?
  - Why should we care?

- What is this class about?
  - What to expect?
  - Logistics
What are we here to discuss?

Cutting-edge advances made in Molecular Modeling using the power of AI
Demo time...

https://www.youtube.com/watch?v=iUMpm3tYsVE

More details:
https://www.science.org/content/article/breakthrough-2021
Modeling life...the inception?

"Principle of Divergence"
Modeling life... the inception?

1859
Life at the macro scale
Life at the molecular scale

DNA

RNA

Protein
The inner life of the cell

https://www.youtube.com/watch?v=wJyUtbn0O5Y

— BioVisions@Harvard
The role of AI

AI
ML
DL
Deep Learning Revolution in Molecular Modeling

- **Progress in deep learning**
  - Deep fully connected NNs
  - ConvNet
  - RNN
  - Graph convolutional NNs
  - ResNet
  - GAN
  - VAE
  - …

- **Molecular data types**
  - Structured data
  - 1D sequence data
  - 2D image or profiling data
  - Graph data
  - 3D coordinate data
  - 4D dynamics data
  - …

- **The post-genomic "big data"**
  - High-throughput DNA sequencing
  - Post Moore's Law Computing
Few examples

<table>
<thead>
<tr>
<th>Example</th>
<th>Model</th>
<th>Data type</th>
<th>Research direction</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enzyme function prediction</td>
<td>DNN</td>
<td>Structured</td>
<td>Biomolecular function prediction</td>
<td>Classification</td>
</tr>
<tr>
<td>Gene expression regression</td>
<td>DNN</td>
<td>Structured</td>
<td>Biomolecular property prediction</td>
<td>Regression</td>
</tr>
<tr>
<td>RNA-protein binding sites prediction</td>
<td>CNN</td>
<td>1D data</td>
<td>Sequence analysis</td>
<td>Classification</td>
</tr>
<tr>
<td>DNA sequence function prediction</td>
<td>CNN, RNN</td>
<td>1D data</td>
<td>Sequence analysis</td>
<td>Classification</td>
</tr>
<tr>
<td>Biomedical image classification</td>
<td>ResNet</td>
<td>2D data</td>
<td>Biomedical image processing</td>
<td>Classification</td>
</tr>
<tr>
<td>Protein interaction prediction</td>
<td>GCN</td>
<td>Graph</td>
<td>Biomolecule interaction prediction</td>
<td>Classification</td>
</tr>
<tr>
<td>Biology image super-resolution</td>
<td>GAN</td>
<td>2D image</td>
<td>Structure reconstruction</td>
<td>Data generation</td>
</tr>
<tr>
<td>Gene expression data embedding</td>
<td>VAE</td>
<td>2D data</td>
<td>Systems biology</td>
<td>DR, Data generation</td>
</tr>
</tbody>
</table>
Shallow neural network

A

Node inputs
\[ x_1 : 1.0 \rightarrow w_1 : -0.32 \]
\[ x_2 : 0.8 \rightarrow w_2 : 0.11 \]

Weights
\[ \sum \]
Bias
\[ 0.002 \]

Activate function \( \delta(z) \)
\[ \max\{0, -0.23\} \]

Node output
\[ 0.00 \]

B

Input layer
Hidden layer
Output layer

Cross-entropy

SGD: update weights

Error backpropagation

Ground Truth: \( \hat{y} \)

Forward propagation

\[ L = -\sum_i \hat{y}_i \log(y_i) \]
\[ = -1.0 \times \log(0.74) - 0.0 \times \log(0.26) \]
\[ = 1.31 \]

Convolution neural network

Recurrent neural network

\[ y = \tanh(w_h \cdot s_5) \]
\[ s_t = \tanh(w_x \cdot x_t + w_{rec} \cdot s_{t-1}) \]

RNN

Variants of convolutional neural networks

Graph neural network

\[ h_{a,2} = w_2 \cdot x_{\text{average}(b,c,d),1} + s_2 \cdot x_{a,1} \]

Generative adversarial network

Autoencoder

Issues to think about

- Lack of data
- Overfitting
- Imbalanced data
- Interpretability
- Uncertainty scaling
- Catastrophic forgetting
- Reducing computational requirement and model compression
“I can’t be as confident about computer science as I can about biology. Biology easily has 500 years of exciting problems to work on. It’s at that level.”

Donald Knuth
What this class is about?

- Surveying the emerging field of computational modeling of molecules with a particular focus on predictive modeling driven by advances in AI and Machine (Deep) Learning.

- **Goal**
  - After finishing this course, you should be ready with a paper to submit to a conference/journal

- **Target audience**
  - PhD students, MS Thesis students (Intro time!)
What this class is NOT

- **NOT the target audience**
  - Coursework-only grad students
  - Students looking to add an AI course to their resume

- **NOT the goal**
  - Teaching a toolkit (e.g., TensorFlow/PyTorch)
  - Programming
Prerequisites

- This course is appropriate for graduate students in computer science, computational biology, bioinformatics, and statistics.
- Familiarity with fundamental concepts in machine learning, statistics, probability and algorithms is expected.
Course Information

- **Course website**
  - [https://people.cs.vt.edu/dbhattacharya/courses/cs6824/](https://people.cs.vt.edu/dbhattacharya/courses/cs6824/)

- **Piazza**
  - [https://piazza.com/vt/spring2022/cs6824/home](https://piazza.com/vt/spring2022/cs6824/home)

- **Canvas**
  - [https://canvas.vt.edu/courses/145337](https://canvas.vt.edu/courses/145337)
Course Staff

- **Instructor**
  - Debswapna Bhattacharya
  - Office Hours: Monday and Wednesday 4:00 pm - 5:00 pm at Torgersen 2160N
Grading

- **Class participation: 20%**
  - Involvement in class - 5%
  - Peer review - 15%
- **Paper presentation: 30%**
  - Each student presents 2 papers (1 before and 1 after spring break each 15%)
  - List of papers will be available for students to pick
- **Project: 50%**
  - Proposal - 20% (10% for whitepaper and 10% for presentation)
  - Report/paper - 15%
  - Presentation - 15%

**Grading scale**
(after computing ceiling of the final percentage of points earned)

- A: 93%-100%
- A-: 90%-92%
- B+: 87%-89%
- B: 83%-86%
- B-: 80%-82%
- C+: 77%-79%
- C: 73%-76%
- C-: 70%-72%
- D+: 67%-69%
- D: 63%-66%
- D-: 60%-62%
- F: Below 60%
Tentative Course Schedule

- **Introductory lectures**
  - Crash course on molecules
  - A general Overview of AI-powered Molecular Modeling

- **Paper presentations**
  - Student presentation of research papers
  - Peer review of papers by students

- **Course projects**
  - Proposal whitepaper and presentation, peer review
  - Final presentation + report
Paper Presentations

- Each student will present 2 papers picked by the student
  - The goal of the presentation is to facilitate a discussion, focusing on:
    - Present the biological question and the corresponding computational abstraction
    - How did the authors address the problem?
    - Did they manage to answer the original biological question?
    - How can we improve the results? What are future directions?

- The remaining students are required to write a short peer review
  - Summary of the paper
  - Major and minor comments
  - Outlook/future directions
<table>
<thead>
<tr>
<th>#</th>
<th>Title</th>
<th>Link</th>
<th>Published In</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Graph Convolutional Policy Network for Goal-Directed Molecular Graph Generation</td>
<td><a href="https://papers.nips.cc/paper/2018/hash/d0d6d78e8f2ba0c54079b8ebbe31117f8-Abs.pdf">https://papers.nips.cc/paper/2018/hash/d0d6d78e8f2ba0c54079b8ebbe31117f8-Abs.pdf</a></td>
<td>NeurIPS 2018</td>
</tr>
<tr>
<td>5</td>
<td>Generative modeling for protein structures</td>
<td><a href="https://papers.nips.cc/paper/2018/hash/5af2b94a1d85c52e75d8a24c3ce53d-Abs.pdf">https://papers.nips.cc/paper/2018/hash/5af2b94a1d85c52e75d8a24c3ce53d-Abs.pdf</a></td>
<td>NeurIPS 2018</td>
</tr>
<tr>
<td>9</td>
<td>End-to-End Learning on 3D Protein Structure for Interface Prediction</td>
<td><a href="https://papers.nips.cc/paper/2019/hash/fc77d7d7f79d61a9c7d1fdffbe05c55-Abs.pdf">https://papers.nips.cc/paper/2019/hash/fc77d7d7f79d61a9c7d1fdffbe05c55-Abs.pdf</a></td>
<td>NeurIPS 2019</td>
</tr>
<tr>
<td>10</td>
<td>Generative Models for Graph-Based Protein Design</td>
<td><a href="https://papers.nips.cc/paper/2019/hash/3a4f4f3396d56a5f460c88caca36d2-Abs.pdf">https://papers.nips.cc/paper/2019/hash/3a4f4f3396d56a5f460c88caca36d2-Abs.pdf</a></td>
<td>NeurIPS 2019</td>
</tr>
<tr>
<td>11</td>
<td>Learning Protein Structure with a Differentiable Simulator</td>
<td><a href="https://openreview.net/forum?id=Bj9y9v3CEKm">https://openreview.net/forum?id=Bj9y9v3CEKm</a></td>
<td>ICLR 2019</td>
</tr>
<tr>
<td>12</td>
<td>Guiding Deep Molecular Optimization with Genetic Exploration</td>
<td><a href="https://papers.nips.cc/paper/2020/hash/8ba5e567b03f7c78d4d495de4f50-Abs.pdf">https://papers.nips.cc/paper/2020/hash/8ba5e567b03f7c78d4d495de4f50-Abs.pdf</a></td>
<td>NeurIPS 2020</td>
</tr>
<tr>
<td>13</td>
<td>Self-Supervised Graph Transformer on Large-Scale Molecular Data</td>
<td><a href="https://papers.nips.cc/paper/2020/hash/54ae58341ae3803abed33a19d5d-Abs.pdf">https://papers.nips.cc/paper/2020/hash/54ae58341ae3803abed33a19d5d-Abs.pdf</a></td>
<td>NeurIPS 2020</td>
</tr>
<tr>
<td>14</td>
<td>Barking up the right tree: an approach to search over molecule synthesis DAGs</td>
<td><a href="https://papers.nips.cc/paper/2020/hash/4c05b352f8375b9f6e7f1d3698-Abs.pdf">https://papers.nips.cc/paper/2020/hash/4c05b352f8375b9f6e7f1d3698-Abs.pdf</a></td>
<td>NeurIPS 2020</td>
</tr>
<tr>
<td>15</td>
<td>Reinforced Molecular Optimization with Neighborhood-Controlled Grammars</td>
<td><a href="https://papers.nips.cc/paper/2020/hash/526bd8fb0be444e0868a2270788-Abs.pdf">https://papers.nips.cc/paper/2020/hash/526bd8fb0be444e0868a2270788-Abs.pdf</a></td>
<td>NeurIPS 2020</td>
</tr>
<tr>
<td>17</td>
<td>Energy-based models for atomic-resolution protein conformations</td>
<td><a href="https://openreview.net/forum?id=S1e_9xqV5S">https://openreview.net/forum?id=S1e_9xqV5S</a></td>
<td>ICLR 2020</td>
</tr>
<tr>
<td>18</td>
<td>Co-evolution Transformer for Protein Contact Prediction</td>
<td><a href="https://papers.nips.cc/paper/2021/hash/708e44bd07586afbf7bb596985-Abs.pdf">https://papers.nips.cc/paper/2021/hash/708e44bd07586afbf7bb596985-Abs.pdf</a></td>
<td>NeurIPS 2021</td>
</tr>
<tr>
<td>20</td>
<td>Language models enable zero-shot prediction of the effects of mutations on protein function</td>
<td><a href="https://papers.nips.cc/paper/2021/hash/5f3387d3695dd42279260780-Abs.pdf">https://papers.nips.cc/paper/2021/hash/5f3387d3695dd42279260780-Abs.pdf</a></td>
<td>NeurIPS 2021</td>
</tr>
<tr>
<td>22</td>
<td>Neural Distance Embeddings for Biological Sequences</td>
<td><a href="https://papers.nips.cc/paper/2021/hash/9a1de01693e05d2551ecbb7e86d8-Abs.pdf">https://papers.nips.cc/paper/2021/hash/9a1de01693e05d2551ecbb7e86d8-Abs.pdf</a></td>
<td>NeurIPS 2021</td>
</tr>
<tr>
<td>26</td>
<td>Learning from Protein Structure with Geometric Vector Perceptrons</td>
<td><a href="https://openreview.net/forum?id=1YLiJd3xU54">https://openreview.net/forum?id=1YLiJd3xU54</a></td>
<td>ICLR 2021</td>
</tr>
</tbody>
</table>
Paper Presentations

- Each student will present 2 papers picked by the student
  - The goal of the presentation is to facilitate a discussion, focusing on:
    - Present the biological question and the corresponding computational abstraction
    - How did the authors address the problem?
    - Did they manage to answer the original biological question?
    - How can we improve the results? What are future directions?

- The remaining students are required to write a short peer review
  - Summary of the paper
  - Major and minor comments
  - Outlook/future directions

Peer-reviews via EasyChair
Course Project

- Can be done individually or in a group (2 students max / project)
- First write and present a proposal, which will receive feedback from fellow students via EasyChair
- Then, conduct research and write a paper
- Pick a venue (conference/journal) and use NeurIPS LaTeX template style for your paper
- Extra credit for shooting for conference/journal submission
Deadlines

Note: All deadlines are until 11:59 PM EST unless otherwise specified

- **Jan 31**: Pick 2 papers to present from [this list](#) and electronically submit the papers to [EasyChair](#) as the corresponding author (publicly discuss your preference in Piazza)
- **Feb 4**: Accept invitation to be a PC member
- **Feb 4**: Submit presentation and peer-review reports no later than 24 hours before each paper presentation
- **Mar 4**: Submit project proposal whitepaper to [Canvas](#) and concurrently upload to [EasyChair](#)
- **Mar 13**: Submit project proposal presentation to [Canvas](#)
- **Mar 21**: Submit proposal peer reviews via [EasyChair](#)
- **Apr 13**: Submit project presentation to [Canvas](#)
- **May 4**: Submit final project report to [Canvas](#) and concurrently upload to [EasyChair](#)
How to stay in touch?

- **Primary means of communication — Piazza**
  - No direct email to instructor unless private information
  - Instructor can provide answers to everyone on forum
  - Class participation credit for answering questions on forum!

- **Class Mailing List**
  - class-cs-6824-20577-202201-g@vt.edu
Policies

- **Collaboration Policy**
  - You are encouraged to collaborate
  - Give proper credit when it's due
  - Project proposal/report will be plagiarism checked

- **Academic integrity**
  - Students enrolled in this course are responsible for abiding by the Honor Code
  - Zero-tolerance philosophy regarding plagiarism or other forms of cheating

- **Principles of Community**
  - The course will include in-class discussions, and we will adhere to Virginia Tech Principles of Community.

- **Accessibility**
  - If any student needs special accommodations because of any disabilities, please contact the instructor during the first week of classes.
  - Such students are encouraged to work with The Office of Services for Students with Disabilities to help coordinate accessibility arrangements.

- **COVID-19 Policy**
  - Please follow the instructions posted at the University and public health guidelines for the latest COVID-19 Policy.
Todo: before next class

Go through the course webpage at:
https://people.cs.vt.edu/dbhattacharya/courses/cs6824/

...and ask any questions in the next class.

Get into Piazza:
https://piazza.com/vt/spring2022/cs6824/home