











# CSC Spring School on Computational Chemistry 2024

This is a collaborative “notebook” for the CSC Spring School in Computational Chemistry organised 17-19.4.2024 by CSC - IT Center for Science.

 The latest updates on the **course schedule** can be found in here.  
 This is also the place to ask questions about the school content!

 **Hint:** HedgeDoc is great for sharing information in this kind of courses, as the code formatting is nice & easy with Markdown (<https://www.markdownguide.org/basic-syntax/>)! Just add backticks ( ` ` ) for the `code blocks` .  
Otherwise, it's like Google docs as it allows simultaneous editing. There's a section for practice down there [↓](#)

- CSC Spring School on Computational Chemistry 2024
  -  How we work
    - Code of conduct
    - Using CSC supercomputers (useful links)
    - Links and info about hands-on materials
  -  Agenda
    - Day 1: Wednesday 17.4.
    - Day 2: Thursday 18.4.
    - Day 3: Friday 19.4.
  -  ICE BREAKER (HedgeDoc -practice)
    - Ice breakers
  -  Q & A
  -  Add your questions here

## How we work

### Code of conduct

#### ▼ Details

We strive to follow the Code of Conduct developed by The Carpentries organisation ([https://docs.carpentries.org/topic\\_folders/policies/code-of-conduct.html](https://docs.carpentries.org/topic_folders/policies/code-of-conduct.html)) to foster a welcoming environment for everyone. In short:

- Use welcoming and inclusive language
- Be respectful of different viewpoints and experiences
- Gracefully accept constructive criticism
- Focus on what is best for the community
- Show courtesy and respect towards other community members

## Using CSC supercomputers (useful links)

### ▼ Details

- Docs CSC (CSC user guides) (<https://docs.csc.fi>)
- Connecting to CSC supercomputers (<https://csc-training.github.io/csc-env-eff/hands-on/connecting/ssh-puhti.html>)
- Puhti web interface (<https://www.puhti.csc.fi>)
- Linux 101 (<https://docs.csc.fi/support/tutorials/env-guide/>)
- Running serial ([https://csc-training.github.io/csc-env-eff/hands-on/batch\\_jobs/serial.html](https://csc-training.github.io/csc-env-eff/hands-on/batch_jobs/serial.html)), parallel ([https://csc-training.github.io/csc-env-eff/hands-on/batch\\_jobs/parallel.html](https://csc-training.github.io/csc-env-eff/hands-on/batch_jobs/parallel.html)) and interactive ([https://csc-training.github.io/csc-env-eff/hands-on/batch\\_jobs/interactive.html](https://csc-training.github.io/csc-env-eff/hands-on/batch_jobs/interactive.html)) batch jobs
- Moving files between local machine and CSC supercomputers using `scp` (<https://docs.csc.fi/data/moving/scp/>) and Puhti web interface (<https://docs.csc.fi/data/moving/web-interface/>)

## Links and info about hands-on materials

### Wednesday 17.4

- Lecture slides: [https://gitlab.com/hseara/csc\\_spring2024\\_notebooks/-/blob/main/MD.pdf](https://gitlab.com/hseara/csc_spring2024_notebooks/-/blob/main/MD.pdf) ([https://gitlab.com/hseara/csc\\_spring2024\\_notebooks/-/blob/main/MD.pdf](https://gitlab.com/hseara/csc_spring2024_notebooks/-/blob/main/MD.pdf))
- MD hands-ons are run using the Mahti web interface, <https://www.mahti.csc.fi> (<https://www.mahti.csc.fi>)
- <https://www.mahti.csc.fi> (<https://www.mahti.csc.fi>) -> Jupyter for courses -> course module `sscc-2024-md`, reservation `sscc_wed_gpu`, partition `gpusmall`
- VMD hands-on is run on Dogmi workstations (or own laptop)
- Materials available at [https://gitlab.com/hseara/csc\\_spring2024\\_notebooks](https://gitlab.com/hseara/csc_spring2024_notebooks) ([https://gitlab.com/hseara/csc\\_spring2024\\_notebooks](https://gitlab.com/hseara/csc_spring2024_notebooks))

### Thursday 18.4

- Lecture slides: [https://a3s.fi/sscc/SSCC\\_2024\\_Quantum\\_Chemistry-2024-04-18.pdf](https://a3s.fi/sscc/SSCC_2024_Quantum_Chemistry-2024-04-18.pdf) ([https://a3s.fi/sscc/SSCC\\_2024\\_Quantum\\_Chemistry-2024-04-18.pdf](https://a3s.fi/sscc/SSCC_2024_Quantum_Chemistry-2024-04-18.pdf))
- **QC basic and QC advanced**
  - Hands-ons are run using TmoleX on local workstation (Dogmi or own laptop) OR <https://www.puhti.csc.fi> (<https://www.puhti.csc.fi>)
  - Slurm advance reservation name: `sscc_thu_small`
  - Partition name: `small`
  - Basic materials available at <https://github.com/csc-training/sscc-qc-basic> (<https://github.com/csc-training/sscc-qc-basic>)
  - Advanced materials available at <https://a3s.fi/sscc/QC-advanced-2024-04-18.pdf> (<https://a3s.fi/sscc/QC-advanced-2024-04-18.pdf>)
- **QC intermediate**
  - Hands-ons are run using the Puhti web interface, <https://www.puhti.csc.fi> (<https://www.puhti.csc.fi>)
  - <https://www.puhti.csc.fi> (<https://www.puhti.csc.fi>) -> Jupyter for courses -> course module `sscc-2024-cp2k`, reservation `sscc_thu_small`, partition `small`
  - Batch jobs use reservation name `sscc_thu_large`, partition name `large`
  - Materials available at <https://github.com/csc-training/sscc-cp2k> (<https://github.com/csc-training/sscc-cp2k>)

### Friday 19.4

- **Enhanced sampling techniques**
  - Lecture slides: [https://gitlab.com/hseara/csc\\_spring2024\\_notebooks/-/blob/main/Enhanced\\_Sampling.pdf](https://gitlab.com/hseara/csc_spring2024_notebooks/-/blob/main/Enhanced_Sampling.pdf) ([https://gitlab.com/hseara/csc\\_spring2024\\_notebooks/-/blob/main/Enhanced\\_Sampling.pdf](https://gitlab.com/hseara/csc_spring2024_notebooks/-/blob/main/Enhanced_Sampling.pdf))
  - Hands-on run using Mahti web interface, <https://www.mahti.csc.fi> (<https://www.mahti.csc.fi>)

- <https://www.mahti.csc.fi> (<https://www.mahti.csc.fi>) -> Jupyter for courses -> course module `sscc-2024-md`, reservation `sscc_fri_gpu`, partition `gpusmall`
- Materials available at [https://gitlab.com/hseara/csc\\_spring2024\\_notebooks](https://gitlab.com/hseara/csc_spring2024_notebooks) ([https://gitlab.com/hseara/csc\\_spring2024\\_notebooks](https://gitlab.com/hseara/csc_spring2024_notebooks))
- **AI for spectroscopy**
  - Hands-on run using Mahti web interface, <https://www.mahti.csc.fi> (<https://www.mahti.csc.fi>)
  - <https://www.mahti.csc.fi> (<https://www.mahti.csc.fi>) -> Jupyter for courses -> course module `sscc-2024-ai4spec`, reservation `sscc_fri_gpu`, partition `gpusmall`
  - Materials available at <https://github.com/csc-training/AI4Spec> (<https://github.com/csc-training/AI4Spec>)
- **Principal component analysis for MD**
  - Lecture slides: [https://gitlab.com/hseara/csc\\_spring2024\\_notebooks/-/blob/main/PCA\\_for\\_MD/ML\\_for\\_chemistry.pdf](https://gitlab.com/hseara/csc_spring2024_notebooks/-/blob/main/PCA_for_MD/ML_for_chemistry.pdf) ([https://gitlab.com/hseara/csc\\_spring2024\\_notebooks/-/blob/main/PCA\\_for\\_MD/ML\\_for\\_chemistry.pdf](https://gitlab.com/hseara/csc_spring2024_notebooks/-/blob/main/PCA_for_MD/ML_for_chemistry.pdf))
  - Hands-on run using Puhti web interface, <https://www.puhti.csc.fi> (<https://www.puhti.csc.fi>)
  - <https://www.mahti.csc.fi> (<https://www.mahti.csc.fi>) -> Jupyter for courses -> course module `sscc-2024-pca`, reservation `sscc_fri_small`, partition `small`
  - Materials available at [https://gitlab.com/hseara/csc\\_spring2024\\_notebooks](https://gitlab.com/hseara/csc_spring2024_notebooks) ([https://gitlab.com/hseara/csc\\_spring2024\\_notebooks](https://gitlab.com/hseara/csc_spring2024_notebooks))

## 17 Agenda

### Day 1: Wednesday 17.4.

Time	Content
8:45	Registration
9:00	Welcome and intro
9:15	<b>Introduction to classical molecular dynamics</b>
10:00	<i>Break</i>
10:15	<b>Introduction to classical MD contd.</b>
11:00	<i>Break</i>
11:15	<b>Practical info and MD hands-on</b>
12:00	<i>Lunch break</i>
13:00	<b>MD hands-on contd.</b>
14:30	<i>Break</i>
14:45	<b>MD hands-on contd.</b>
16:15	<i>Break</i>
16:30	<b>Poster session</b> with snacks, refreshments
19:00	Poster session finishes

### Day 2: Thursday 18.4.

Time	Content
9:00	Welcome and intro
9:10	<b>Introduction to quantum chemistry</b>
10:00	<i>Break</i>
10:15	<b>Introduction to QC contd.</b>
11:00	<i>Break</i>
11:15	<b>Introduction to QC contd.</b>
12:00	<i>Lunch break</i>
13:00	<b>QC hands-on</b>
14:30	<i>Break</i>
14:45	<b>QC hands-on contd.</b>
16:15	<i>Transition to sauna lobby</i>
16:45	<b>Talk, dinner and sauna</b>

### Day 3: Friday 19.4.

Time	Content
9:00	Welcome and intro
9:10	<b>Enhanced sampling methods</b>
10:00	<i>Break</i>
10:15	<b>Enhanced sampling hands-on</b>
11:00	<i>Break</i>
11:15	<b>Enhanced sampling hands-on contd.</b>
12:00	<i>Lunch break</i>
13:00	<b>AI/ML for molecular dynamics</b>
13:45	<b>AI/ML for quantum chemistry</b>
14:30	<i>Break</i>
14:45	<b>AI/ML hands-on (PCA for MD / AI for spectroscopy)</b>
16:15	Closing and goodbyes

## ICE BREAKER (HedgeDoc -practice)

### Ice breakers

Let's learn how to use this HedgeDoc document by answering an ice breaker question!

**Q1: What is your research focused on / what methods have you used in your research/studies so far?**

**Answers:**

- Type here!
- polymer folding, and vesicle fusion driven by dynamic combinatorial chemistry, HPLC, NMR, TEM, SAXS
- Using transformer neural networks for molecular structure optimization

- DFT & MD calculations, multiscale simulation pipeline development, organic molecules & inorganic structures, quantum computing integration and machine learning integration to DFT
- Experimental chemistry with a machine learning aspect regarding catalyst structure and performance.
- MD simulations of lipid-based pharmaceutical formulations
- MD simulations of protein-protein interactions and protein dynamics
- DFT calculations for XPS spectra of atmospheric particles and molecules
- Research is focused on transport of solutes in groundwater. MD can be used to study geochemical and transport processes in complex environments
- MD simulations of ionic liquids for transdermal drug delivery
- Reaction mechanism computations of volatile compound autoxidation in the atmosphere, DFT optimization and WFT corrections with gaussian and molpro
- MD simulations, DFT simulations for XAS/XES/XPS, machine learning
- MD simulations of protein conformational dynamics, protein-protein interactions and membranes. Computational drug discovery and lead optimization
- MD simulations of polymeric systems
- MD simulations of curved lipid membranes and nonaarginine peptides (cell penetrating peptides)
- Experimental chemistry combined with DFT calculations for reaction mechanism studies
- MD simulations of protein variants
- What about AIMD?
- Research subject: Design and study of low-dimensional heterostructures using machine learning and DFT calculations
- Evaluating MD against NMR experiments
- Analyzing how well different forcefields perform on intrinsically disordered proteins
- Automizing selection of quality varified MD data
- MD simulations with non-equilibrium approaches for protein-DNA systems
- DFT calculations for metallocene-methylaluminoxane catalyst for polyolefin synthesis.
- Computational catalysis, DFT calculations
- MD od short cationic peptides
- MD simulations in small molecules


**Q2: What are your expectations for today's sessions?**


- Answer here!
- To get a deeper understanding about QC
- Having fun
- I expect to learn more about the theory behind the practical computations I've used
- I would be interested in X-ray spectroscopy using today's software
- I expect to get some background info on QC
- to learn background and basic theory behind QC.
- How to use the different approximations
- Refurbish h mwy meomuoryl do n
- i would like to find positions at least of few electrons
- Refurbish my memory on the theory and practice of quantum chemistry calculations
- To learn a thing or two :)
- To refresh my skills with quantum chemistry tools
- Refresh the knowledge about quantum chemistry and get some basics of quantum computing
- To learn new things about QC methods
- to get a good start for QC

### Q3: Name one interesting/surprising thing you've learned during the spring school so far!

- Answer here!
- The convenience of Python notebooks for simulation setups. x6
- The cp2k practical was great. :)
- h atom is not regarded as classical objects but quantum.
- how some simple quantum computation can look like
- How wide of a range of applications that computational chemistry has
- the brilliance of people around me x4
- Many new Python libraries
- Many new systems and ways to calculate
- How well this whole course is organized and how easy it is to follow the hands-on exercises x4

## Q & A

Your questions are answered here. We will answer them, and this document will store the answers for you for later use! 

Scroll  to the bottom of the page to submit a question

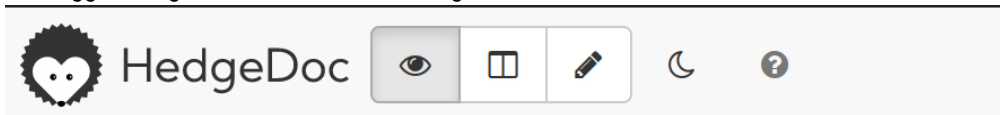
#### I have difficulty pasting my questions into HedgeDoc (here). Do you have some instructions on how to write here?



- A: HedgeDoc can look a bit different depending on the view you are in. Can you see the three icons on top left corner, next to HedgeDoc text? There's pencil, this side-by-side symbol, and an eye (see below for a screenshot). In eye view, you can't edit, you are just viewing. The other two reveal the markdown (MD) version of the page, which you can edit. I find it easiest to edit with the side-by-side view, but it can be a bit slow sometimes. First time opening the page, the small Edit (pencil) button might be on right next to the table of contents. Please note that it might make things faster if you **switch to view only mode when not editing!**

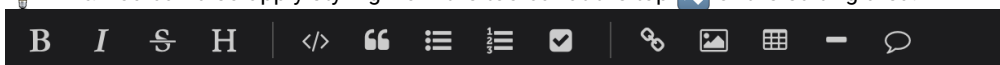
The small pencil icon for entering edit mode for the first time. You might need to scroll all the way up to see it.

16 views  s

The bigger HedgeDoc toolbar for switching between modes later on.




 **Hint:** You can also apply styling from the toolbar at the top  of the editing area.



## Add your questions here

Please type your questions ON THE BOTTOM OF THIS SECTION. We will answer them, and organise the document topically.

- ✓ **Q0: Have I clicked the edit mode on?**
- A: Probably not yet... 
- ✓ **Q1: Hector mentioned a GitHub with the materials. Could we have the link? Would be easier to also follow the slides on our own screens.**
- A: Yes, you can find the materials here: [https://gitlab.com/hseara/csc\\_spring2024\\_notebooks](https://gitlab.com/hseara/csc_spring2024_notebooks) ([https://gitlab.com/hseara/csc\\_spring2024\\_notebooks](https://gitlab.com/hseara/csc_spring2024_notebooks)). Link also at the beginning of this page
- ✓ **Q2: I missed the part where the constants (ks) used in the intramolecular interactions were obtained. How do you get these ks??**
- A: Spectroscopy is one way (so experimental data), or then you can get them by fitting to accurate ab initio (quantum chemical) calculations.
- ✓ **Q3: Thanks, but then we could use machine learning to reduce the amount of points to calculate by ab initio right? (this question can be asked later)?**
- A: Indeed, ML methods are nowadays frequently used to train machine learned interatomic potentials that can be as efficient as classical forcefields but (almost) as accurate as ab initio.
  - One issue is that the training data for biomolecules (apart from proteins used in, e.g., AlphaFold) is scarce. But it seems that ML-based force fields trained on some small molecules/fragments perform reasonably well for at least lipids.
- ✓ **Q5: Does anybody here do AIMD and machine learning to train machine learning potentials? I'd like to talk to someone about it to gain some potential tips.**
- A: Antti has some experience here!
- ✓ **Q6: Are there many benchmark papers nowadays to find issues of FFs? Is not the goal to sell best (and hide issues) nowadays?**
- A: It is true that issues are being swept under the carpet frequently, but that is not good science. Luckily, there are many review papers around and plenty of researchers who still publish "failure" papers where things have not worked out e.g. due to force field issues. We encourage you to do the same and contribute to openness in science!
- ✓ **Q7: Related to the protein-membrane tutorial: What is the easiest way of calculating cholesterol tilt with MDAnalysis?**
- A: A simple solution would be to use the approach mentioned in this paper (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2919319/>). Create an index group with two cholesterol atoms (C3 and C17). Then, using gromacs tool 'angle', calculate the angle (using the flag '-type angle') between the vector joining these two atoms and the membrane normal.
- ✓ **Q8: Is there an actual difference between ab initio and first principles methods?**
- A: These two are often taken as synonymous terms for the same thing, i.e. methods that have no system-dependent parametrization. Literally, ab initio means "from the beginning".
  - Sometimes ab initio is used for wavefunction based methods while first-principles may also include DFT (which often has a little bit of empirical parametrization e.g. in the exchange-correlation functionals).
- ✓ **Q9: Are these slides (Thursday) available somewhere?**
- A: Links to all materials are available at the beginning of this page
- ✓ **Q10: Why is the kinetic energy a negative term in the expanded Schrodinger's equation?**
- A: When dealing with the total energy of a many-electron system, lower energy (more negative) means a more stable system (closer to ground state). A higher kinetic energy

contributes to lowering the energy. However, one needs to remember that the kinetic energy of electrons is a quantum mechanical concept, so it is a bit hard to interpret it intuitively as something similar to classical kinetic energy.

✓ **Q11: You mention the mean field approximation as something that could improved on, does it mean that mean field approximations can be considered obsolete or can they still find use in nowadays research?**

- A: Mean-field approximations are very relevant still, as they serve as our baseline when constructing more accurate approximations. E.g. post-HF theories build upon the mean-field approximation and DFT is also in some ways a mean-field theory.

✓ **Q12: Regarding the polarization functions, from the example with p and d orbitals, it looks like an hybridization such as sp<sup>2</sup>, sp<sup>3</sup> carbons. Are polarization and hybridization synonyms here?**

- A: No. Hybridization is a much simpler concept that was invented to describe valence orbitals qualitatively. They may still be sometimes useful e.g. in organic chemistry. More generally, however, molecular orbital theory is a more accurate theory. The resemblance is there because ultimately we are describing the same thing (molecular orbitals), but in proper MO theory we are not confined to the simple hybridization "rules".

✓ **Q13: Is there a rule of thumb for when polarization functions in basis sets are expected to significantly affect the results?**

- A: Polarization functions are important when we want to describe orbitals that are asymmetric around the nuclei. So these are basically very important for most systems of practical interest. E.g. to describe chemical bonding, polarization functions are a must since chemical bonds are typically polarized.

✓ **Q14: You mentioned something about "all-electron" implementations. Can you elaborate more on this term and what "types" of implementations are available for quantum chemistry calculations?**

- A: All-electron means that also the core-electrons (not just valence electrons) are treated explicitly. Often one may alternatively use effective-core potentials (ECP), aka. pseudopotentials. These potentials condense the contributions of the core electrons into a term that is not solved as part of the SCF cycle (only valence electrons are included). They often also include relativistic effects which are relevant for heavy elements.

✓ **Q15: What exactly does "r" represent in the density function? Coordinates over all space, or specific coordinates? Individual electrons?**

- A:  $\mathbf{r}$  (or  $\vec{r}$ ) corresponds to electronic coordinates in 3D space. For a single-electron system it is the position of that electron, for many-electron system it contains the positions of all electrons. For  $N$ -electron system notation  $\mathbf{r}^N$  is often used.

✓ **Q16: Can machine learning help in conceiving better functionals? If yes how?**

- A: It can be done, but thus far it has not produced any breakthrough functionals. One example is BEEF-vdW functional.

✓ **Q17: Is it advised to "play" with functionals, i.e. using one for geometry optimization and another one for the subsequent steps?**

- A: This could be viable approach e.g. if geometry optimization using a computationally heavy functional is too costly. So converge the geometry using a lower-level functional (e.g. GGA) and then just reconverge the energy (electronic structure) at a higher level of theory.
- A: one should be mindful that many XC functionals perform very well due to error cancellation. So by being an approximation, there are some effects that may be underestimated due to some simplification of the functional, but on the other hand other simplifications have a similar overestimating effect -> you get quite good results (but for



wrong reasons). If you start to mix different functionals, you might actually get worse results since you might lose this kind of fortuitous error cancellation.

✓ **Q18: Talking about the use of symmetry, while making the calculations faster can they also make shortcuts that could be eventually harmful for the simulation?**

- A: Sure, if there e.g. are some conformational degrees of freedom (i.e. the molecular symmetry), constraining the calculation to a specific point group may have unwanted effects.

✓ **Q19: Does it make sense to split a calculation into one super cheap calculation that gets you most of the way to the solution, and then another one to take it to the proper end? For example, use a cheap functional first, and then a much more sophisticated one.**

- A: Sometimes, yes! E.g. in the CP2K hands-on we performed a NEB calculation to optimize a reaction path at GGA-level, and then reconverged just the energies using a more costly hybrid functional to get an improved estimate of the energy barrier. See also the answer to question Q17.

✓ **Q20: Is there a link for today's presentation? Couldn't find it at the top.**

- A: It is in the same gitlab repository as the MD lecture slides: [https://gitlab.com/hseara/csc\\_spring2024\\_notebooks](https://gitlab.com/hseara/csc_spring2024_notebooks) ([https://gitlab.com/hseara/csc\\_spring2024\\_notebooks](https://gitlab.com/hseara/csc_spring2024_notebooks))

✓ **Q21: What are examples of nonergodic systems/events?**

- A: Experimentally, these could be some form of confined systems. A system that is trapped in a particular state and cannot escape it. Example of a blinking quantum dot in a salt was given where not average of blinking time exists as it depends on the measurement time. The longer you measure the longer spacings between the blinkings appear. On the practical side because MD simulations are short respect to the experimental counter parts, our systems can be considered "non-ergodic" and may require advanced sampling methods to avoid such computational limitation.

✓ **Q22: Would it be possible to use replica exchange somehow to form a phase diagram for a system? If yes, should the temperatures near e.g. the main transition temperature be sampled more rigorously with smaller temperature steps?**

- A: To some extent yes. As you will have the proper sampling for each of the temperatures of your windows. However, because we must ensure exchange between windows, we are often limited on the range of temperatures as many windows are needed to create near-equilibrium jumps.

✓ **Q23: In replica exchange, how is "close enough" defined to justify the exchange?**

- A: In replica exchange we attempt exchanges every  $n$  steps. However, we do not force them. We check if the potential energy of the two frames we want to exchange are close,  $U(T1)$  and  $U(T2)$ . Then you can decide using for example a metropolis criterion.  
 $P(\text{swap}) = \min(1, \exp(-\Delta))$

✓ **Q24: How high can we get in terms of temperature to sample? Are we not limited by the stability of molecules?**

- A: You can get as high as you want, but you will increase dramatically the number of windows that you will use making unfeasible the method. Plus you should take into consideration that you may impact the molecular stability of some of the components such as proteins. Otherwise, no limits, MD allows you to go really wild.

✓ **Q25: What is the main reason umbrella sampling has been so popular and Jarzynski/Crooks not? What is their main difference in practice?**

- A: I will reply in practical terms not really reflecting the theory. In my experience, Non equilibrium methods look awesome in paper, but when you try to implement them you require so many realizations (replicas) that become more expensive than the equilibrium

sampling methods discussed. Also, we can reuse the biased equilibrium simulations to extract properties of the systems using reweighting techniques. This is not really feasible with non-equilibrium methods. Finally, the Jarzynski/Crooks methods only work in the linear regime close to the equilibrium is this is sometime not easy to asses, you must have gaussian distributions to check the overlap, otherwise ...

✓ **Q26: Are these ML slides somewhere? They weren't at the top nor in the github.**

- A: They are in the gitlab, under folder "PCA for MD", [https://gitlab.com/hseara/csc\\_spring2024\\_notebooks/-/tree/main/PCA\\_for\\_MD](https://gitlab.com/hseara/csc_spring2024_notebooks/-/tree/main/PCA_for_MD) ([https://gitlab.com/hseara/csc\\_spring2024\\_notebooks/-/tree/main/PCA\\_for\\_MD](https://gitlab.com/hseara/csc_spring2024_notebooks/-/tree/main/PCA_for_MD))

✓ **Q27: How much model interpretability coming from the features is retained during principal component regression? Since the data dimension is reduced to its principal components, do these principal components retain any physical meaning?**

- A: As long as the base features you chose for the Dimensionality reduction (DR) with PCA are meaningfully associated with your choice of interpretation, the coefficients of these features are indeed meaningful (with a small caveat; you have to normalize your data first). Comparing the coefficients gives you an idea which features are highly important in each component aka mode of the PCA.

✓ **Q28: Nowadays, NN are very famous, have a lot of hype. Are they the answer to everything though?**

- A: Haha. A bit subjective answer to this question has to do with two concepts. Interpretability and Explainability. A model is explainable if you understand how the architecture of the model transforms the input into an output. Most modern NN algorithms are quite explainable as smart people have created architectures that apply explainable transformations. Interpretable models requires that you understand what transformation was actually applied and why it produced the result. This latter part is quite hard with NN based methods. So NN methods are very powerful for prediction tasks and segmentation tasks. However, they, as a rule, can not provide mechanistic insight without specialized architectures.

✓ **Q29: If the input is known, but the algorithm less (not to a mathematical level), are we still having a GIGO problem? GO only maybe?**

- A: Not really. the input in a typical ML problem is not just the raw data but also its "featurization", i.e. selecting how you feed it to the model. Some models require a featurization with certain properties, for example if your data is not Normally distributed should you be using certain statistical tests? Similar constraints might also prevail for ML models. As a result understanding the requirements of the model might determine the optimal way to feed data to the model might be required to not to create GI outcome.

✓ **Q30: Is knowing the model perfectly preventing from analyzing and rationalizing the outputs, therefore limiting the GO problem? ?**

- A: if I understand the question correctly, then indeed knowing the model well helps you avoid using it for the wrong interpretation. For example. Dimensionality Reduction methods have certain expectations about the data generation process (e.g. Diffusion based DR requires that your data was at least generated from a continuous walk process. If you use disjoint data, i.e. data generated from different parts of disconnected space, these models might not yield a meaningful DR.)

✓ **Q31: Is there a need for data cleaning before or after the representation step?**

- A: Sometimes yes. Outlier detection and removal might be important if your methods are very easily misled by extreme values which to your model building intuition appear as artefacts.

✓ **Q32: Spectra also has a y-dimension for intensity (typically). How were these derived from data points that contained only energies (x-dimension)?**

- A: Each eigenvalue was first positioned as a delta-function on the x-axis. Then each delta-function was expanded with a Gaussian function with a certain spread (like 0.1 eV, it's in the publication). The result was a bunch of overlapping Gaussians of the same height. Then we summed up all the Gaussians and that produced plausible spectral data, with intensity variations arising from the sum of the signals.

✓ **Q33: Would it be interesting to use only the partial charges around the atoms constituting the molecule for a coulomb-like matrix?**

- A: Nice idea, and some groups have been learning partial charges. But the partial charges change and would have to be computed for each molecule by diagonalising the Hamiltonian, and that is the calculation we are trying to avoid! CM descriptor is used a lot because it can be built just from element and positioning information, no computation needed.

✓ **Q34: How is "average prediction" defined? Is there a scoring function value?**

- A: We evaluated the mean absolute error (MAE) for each spectrum predicted by subtracting the prediction from the real spectrum at each of the 300 x points and summing up all the differences. This gave us a histogram of MAEs, so we looked at the top, bottom and middle for examples of best, worst and average predictions.

✓ **Q35: In the "worst" prediction, how is the negative band explained (higher energies)?**

- A: There is no physical meaning to the small negative prediction of intensity, but we did not set it to zero. It just enters into our error summation. The model is supposed to learn that all spectra are always positive, and it generally does that.