Student Project Proposal

Bayesian Correction of DFT-Based NMR Predictions Using Reference Molecules

Master area	AI for Drug Discovery; computational chemistry; Bayesian modelling; molecular spectroscopy
Project type	Computational chemistry and probabilistic modelling project
Duration	2-3 months
Supervision	One tutor or co-tutor; weekly or bi-weekly progress meetings
Recommended level	Basic Python, statistics and molecular modelling; prior exposure to NMR or DFT is useful but not mandatory
Main output	Documented workflow/notebook, short technical report and final presentation
Possible data	Small reference set of molecules with experimental NMR data and DFT-predicted chemical shifts; tutor-provided or literature-derived data

Short description. The project asks the student to develop a physics-informed Bayesian calibration workflow that corrects DFT-based NMR predictions using a small set of reference molecules and provides uncertainty on the corrected spectrum.

1. Background and rationale

DFT calculations are widely used to predict NMR chemical shifts and, when feasible, spin-spin coupling constants. These predictions support structure elucidation and molecular assignment, but direct agreement with experimental spectra is often affected by systematic bias, conformational averaging, solvent effects and experimental uncertainty.

A purely data-driven model may require more experimental spectra than are realistically available. A more robust strategy is to keep DFT as the primary physical model and use Bayesian statistics to correct the residual error between theory and experiment.

This project is relevant to AI for drug discovery because many discovery workflows require molecular identification, structural confirmation and uncertainty-aware interpretation of computed observables.

2. Project objective

The objective of the project is to implement a Bayesian correction layer for DFT-based NMR predictions. Reference molecules with both experimental and calculated NMR observables will be used to estimate systematic residuals, and the resulting model will be transferred to one or more target molecules.

The project should focus on a manageable dataset and a clear proof of concept. It is acceptable to use precomputed DFT values if new quantum calculations are too time-consuming.

3. Specific aims

- Collect or organize a small set of reference molecules with experimental and DFT-predicted NMR observables.
- Calculate residual errors between experimental and theoretical chemical shifts and, optionally, coupling constants.
- Fit a Bayesian regression or hierarchical calibration model including global, chemical-class and molecule-level effects.
- Apply the calibrated model to target molecules and generate posterior corrected NMR parameters.
- Reconstruct raw and corrected spectra and compare them quantitatively.
- Report uncertainty intervals and discuss when the correction is reliable.

4. Proposed methodology

4.1 Data preparation

The student will create a table containing molecule identifiers, atom labels, DFT-predicted observables, experimental values and chemical environment annotations. Missing values and ambiguous assignments should be documented explicitly.

4.2 Residual analysis

The residual error will be defined as the difference between experimental and calculated values. Initial plots will be used to identify global offsets, scaling effects and environment-dependent deviations.

4.3 Bayesian calibration model

A simple model may express the experimental chemical shift as a linear function of the DFT value plus global, class-specific and molecule-specific terms. The model should return posterior distributions rather than only point estimates.

4.4 Transfer to target molecules

After calibration, the model will be applied to target molecules processed with the same DFT protocol. Corrected chemical shifts will be used to reconstruct a predicted spectrum with uncertainty bands.

4.5 Validation

The student will compare raw DFT and corrected predictions using MAE, RMSE, correlation and visual spectral comparison. Leave-one-molecule-out validation may be used when the reference set is small.

5. Indicative work plan

Period	Main task	Expected output
Weeks 1-2	Reference set definition	Curated table of molecules, atom assignments and available observables.
Weeks 3-4	Raw DFT vs experiment analysis	Residual plots and baseline error metrics.
Weeks 5-6	Bayesian model implementation	Calibration notebook with posterior parameters and diagnostics.
Weeks 7-8	Target molecule correction	Corrected shifts and reconstructed spectra with uncertainty.
Weeks 9-10	Validation and sensitivity analysis	Comparison of raw and corrected predictions; discussion of limits.
Weeks 11-12	Final report and presentation	Final workflow, report and slide presentation.

6. Expected deliverables

- A curated reference dataset with DFT and experimental NMR observables.
- A documented Python notebook implementing the Bayesian correction model.
- Raw and corrected spectra for at least one target molecule.
- Quantitative comparison between uncorrected and corrected predictions.
- A final technical report and a 10-15 minute presentation.

7. Skills acquired by the student

- Integration of quantum chemical predictions with experimental data.
- Bayesian modelling and uncertainty-aware calibration.
- Data curation for molecular spectroscopy.
- Model validation under limited-data conditions.
- Critical interpretation of computed molecular observables.

8. Suggested software and resources

The project can be implemented with pandas, NumPy, scipy, matplotlib and PyMC, Stan or another Bayesian modelling package. RDKit may be used for chemical annotation if needed.

If the DFT calculations are not part of the student task, the tutor should provide calculated chemical shifts and a small set of experimental references.

9. Optional extensions

Possible extensions include separate models for proton and carbon chemical shifts, incorporation of chemical environment descriptors, leave-one-molecule-out validation, or comparison with a simple non-Bayesian linear correction.

10. Evaluation criteria

Criterion	What should be assessed
Scientific clarity	The scientific question, assumptions and expected use case are clearly stated.
Technical implementation	The code is readable, documented and reproducible on the selected data.
AI relevance	The project produces representations, models or workflows that are suitable for AI-based analysis.
Interpretability	The student explains what the results mean in a drug discovery or molecular science context.
Critical discussion	The final report discusses limits, uncertainty and possible future extensions.

Note for tutors. The main learning objective is not to run the largest possible DFT workflow, but to show how a physically meaningful computational prediction can be corrected using limited experimental evidence and Bayesian uncertainty.