ENVIRONMENT-CONDITIONED MOLECULAR SPECTRAL PREDICTION: A CONTROLLED STUDY OF DATA, REPRESENTATION, AND ARCHITECTURE

Anonymous authors
Paper under double-blind review

ABSTRACT

Molecular absorption cross sections govern atmospheric retrieval and biosignature detection, but experimental coverage extends to only a few hundred species, motivating learned emulators. Unlike the single-condition spectra targeted by prior ML spectroscopy work, atmospheric cross sections are environment-conditioned: the same molecule absorbs differently across temperature and pressure. The available labels also carry source-dependent floors. Experimental spectra have an instrumental detection limit near 10^{-30} cm^2/molecule, whereas simulated spectra extend down to 10^{-255}. We formulate this task as learning a molecule-conditioned continuous function over wavenumber and evaluate it on 570 experimental HITRAN species under 4-fold molecule-wise cross-validation, comparing 29 configurations across 6 design axes with record-level paired significance testing on a fixed set of 1493 held-out spectra. Synthetic data augmentation is the biggest contributor to test performance on real experimental data: synthetic-only training does not transfer to measured spectra (median $R^2 = -71.5$), yet 50% synthetic augmentation in training lifts the median R^2 from 0.46 to 0.60 over experimental training alone. Masking sub-threshold simulated points from the loss beats clipping them at the limit, which is indistinguishable from no treatment. Given these choices, simple encodings win or tie with expressive ones, and equivariant backbones give no advantage over a plain invariant GNN at a matched 5M-parameter budget.

1 INTRODUCTION

Molecular spectroscopy plays a fundamental role across a wide range of scientific disciplines, including atmospheric chemistry, materials science, pharmaceutical development, and planetary science (Herzberg & Spinks, 1939; Herzberg, 2008; Bernath, 2005). In this paper, we focus on planetary and exoplanetary atmospheres, where spectra arise from dense forests of rotational-vibrational transitions whose intensities and line shapes depend sensitively on thermodynamic conditions. Temperature determines quantum-state populations through Boltzmann statistics; pressure governs collisional broadening and line mixing. Together, they can reshape spectral signatures by an order of magnitude even within moderate thermodynamic ranges relevant for habitability, as Fig. 1 (*left*) shows for a single species.

Physics-based simulators model these spectra accurately but cover few species and are expensive to run. Databases such as HITRAN (Gordon et al., 2022), HITEMP (Rothman et al., 2010), and ExoMol (Tennyson et al., 2016) provide line-by-line transition parameters that support physically grounded synthesis and form the backbone of atmospheric retrieval pipelines. However, complete line lists exist for only a few dozen species, and measured cross sections for several hundred, while retrieval and biosignature searches require predictions for thousands to tens of thousands of molecules across diverse environments (Seager et al., 2016). In addition, line-by-line radiative transfer calculations are computationally expensive, making it difficult to marginalize over large molecular inventories or perform rapid inverse modeling.

Learned emulators address the cost but have so far targeted a different problem. Neural networks trained on large spectroscopic datasets perform well when spectra are measured or computed at

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

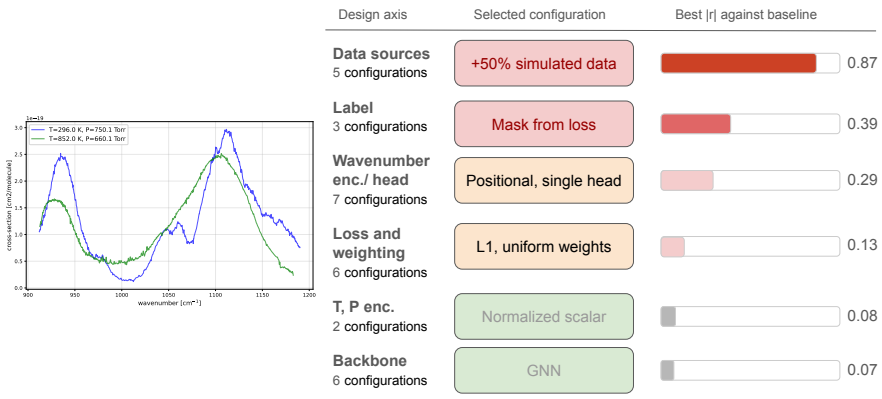


Figure 1: *left* Measured high-resolution absorption spectra of 1,3-dioxolane at two thermodynamic conditions. Temperature and pressure reshape line intensities and profiles, and our goal is to learn this relationship directly from molecular structure and (T, P). *right* We perform a controlled study of the impact of six design axis with 29 configurations, the rank-biserial correlation r is read as small/medium/large at $0.1/0.3/0.5$, showing that training-data composition is the only axis on which any configuration wins by a large effect.

a single standardized condition near room temperature and moderate pressure (Qiao et al., 2020; Batzner et al., 2022). Graph neural networks such as SchNet (Schütt et al., 2017) and GemNet (Gasteiger et al., 2021) map molecular geometry to quantum-mechanical properties, and equivariant message-passing models such as PaiNN (Schütt et al., 2021) extend this to tensorial quantities that downstream simulators convert into spectra. In all of these, the spectrum is treated as an intrinsic molecular property: either predicted at one fixed condition (McGill et al., 2021), or reached indirectly through intermediate quantum-chemical quantities that a physics-based simulator turns into a spectrum (Zou et al., 2023). Atmospheric spectroscopy operates in a different regime, in which the spectrum is an environment-conditioned observable rather than a molecular constant.

We therefore explore what actually matters for a model that predicts cross sections directly from structure and thermodynamic state. We run 29 configurations across six design axes (Fig. 1 *right*) under a fixed evaluation protocol, and find that the synthetic data augmentation and label treatment of the training set are the largest contributors. Our contributions are:

- We formulate the task of environment-conditioned cross section prediction as molecule-conditioned continuous regression over wavenumber, with molecule-wise CV and per-record paired testing.
- A controlled comparison of the impact of different levels of synthetic data augmentation showing that an optimal amount of synthetic data can significantly boost performance for predicting real experimental spectra.
- A controlled comparison of the impact of different treatments of the synthetic label floor, showing that masking simulated entries below the instrumental detection limit from the loss gives a medium boost over no treatment
- A controlled decomposition of design spaces, showing that simple representations and architectures win or tie with expressive ones, and equivariant backbones give no advantage over a plain invariant GNN at a matched 5M-parameter budget.

2 Related Works

Learned spectral emulators. Message-passing GNNs are the standard representation for molecular property prediction, encoding atom types and local neighborhoods while remaining permutation invariant. The same machinery has been applied to spectra: McGill et al. (2021) predict infrared spectra with message-passing networks, and DetaNet predicts several classes of organic molecular

spectra while introducing the QM9S dataset (Zou et al., 2023). Related approaches predict intrinsic quantum-chemical quantities that a physics-based simulator subsequently converts into a spectrum (Qiao et al., 2020). In each case the spectrum is an intrinsic molecular property, computed or measured at a single reference condition. We instead take temperature and pressure as explicit inputs and require generalization across thermodynamic conditions and unseen molecules.

Geometric inductive bias. Molecular observables obey Euclidean symmetries, and a substantial line of work builds architectures that respect them. NequIP shows that fully $E(3)$-equivariant message passing yields large data-efficiency gains for interatomic potentials (Batzner et al., 2022); SEGNN generalizes steerable equivariant message passing to arbitrary geometric quantities via irreducible representations (Brandstetter et al., 2021); EquiformerV2 scales equivariant attention by replacing full tensor products with SO(2) linear layers (Liao et al., 2024); and GotenNet targets the same efficiency question with parameter-efficient tensor interactions (Aykent & Xia, 2025). Equivariant message passing has been connected directly to spectrum-relevant targets (Schütt et al., 2021). A parallel line keeps angular information without steerable tensors: MGNN encodes angles as scalar moments over bond-vector triplets, remaining rotation-invariant rather than equivariant (Chang & Zhu, 2025). Graph transformers replace fixed neighborhood aggregation with learned attention (Shi et al., 2021). We compare representatives of all three families at a matched 5M-parameter budget, and find that none of these inductive biases separates from a plain invariant GNN on this task. We interpret this result narrowly in §5.6, since our geometries are generated conformers and our target is an invariant scalar field.

3 Task and Data

This section defines the prediction target and the two data sources: we state the task (§3.1), the molecular inputs (§3.2), the two datasets (§3.3), and the label-floor mismatch between them that motivates one of our six design axes (§3.4).

3.1 Task formulation

We study the prediction of molecular absorption cross sections as a function of thermodynamic state. A molecule is represented by its atomic graph $\mathcal{G} = (V, E)$, with each atom $i \in V$ carrying an atomic number $Z_i \in \mathbb{N}$ and a position $\mathbf{r}_i \in \mathbb{R}^3$ in Ångström; E contains covalent bonds together with a radius graph (§3.2). Let $\nu \in \mathbb{R}^+$ denote wavenumber (cm^{-1}), $T \in \mathbb{R}^+$ temperature (K), $P \in \mathbb{R}^+$ pressure (atm), and $\iota \in \mathbb{N}$ the isotopologue index. The task is to learn

$$f_\theta : (\mathcal{G}, \nu, T, P, \iota) \;\mapsto\; \log_{10} \sigma(\nu \mid T, P, \iota), \qquad \sigma \; [\mathrm{cm}^2/\mathrm{molecule}]. \tag{1}$$

We work in $\log_{10}$ because σ spans fourteen decades within an experimental record (one record is a single measured spectrum of a species at one (T, P)) and more than two hundred across the simulated dataset. f_θ is a graph encoder producing a molecule embedding, followed by a coordinate decoder queried at (ν, T, P); this separation is what makes T and P decode-time coordinates rather than node features, and lets a single trained model be evaluated on an arbitrary wavenumber grid at test time.

Unlike many molecular property prediction tasks, the output here is a high-dimensional structured signal rather than a single scalar or low-dimensional tensor, and it is not intrinsic to the molecule, as illustrated in Fig. 1(*left*).

3.2 Molecular representation

Atoms carry a 23-dimensional feature vector: 22 electronic-configuration features indexed by atomic number, plus the isotopologue index. Edges are the covalent bonds from the SMILES string, typed by bond order and aromaticity, unioned with a 5 Ångström radius graph so that non-bonded contacts are represented. We generate a single conformer per species with RDKit ETKDG and relax it with MMFF. This single-conformer choice bounds what any geometry-aware architecture can exploit, and we return to it when interpreting the architecture comparison in §5.6.

3.3 DATA SOURCES

We draw on two HITRAN products with complementary coverage and fidelity (Table 1). The **experimental** dataset is the measured absorption cross-section (XSC) archive: broad chemical coverage, real instrument noise, narrow and irregular spectral windows. The **simulated** dataset is computed line-by-line from HITRAN transition parameters with Voigt line shapes on a dense uniform grid: exact by construction, dense in (T, P), but restricted to the few dozen species for which line-by-line lists exist.

Experimental records are retained if they cover at least 10% of the 300-10,000 cm^{-1} analysis window and contain at least 60,000 measured points in it, giving 1493 records over 570 molecules. Because of computation cost, every simulated species is small (≤ 7 atoms); this motivates the size strata used throughout our evaluation (small ≤ 7, medium 8-15, large > 15 atoms). The two datasets are also chemically disjoint, which we enforce explicitly: all 13 isotopologues of the 8 molecules appearing in both datasets are removed from the simulated side globally (in Appendix A), before any fold assignment, so no fold model ever sees a simulated version of a test species.

Simulated spectra are computed on a 0.02 cm^{-1} grid spanning 300-10,000 cm^{-1} (485,001 bins). To bound memory, we subsample each experimental and simulation record to 60,000 points, drawn without replacement from the measured positions with a fixed per-record seed, so every configuration of every comparison is scored on an identical point set. Sampled indices are retained, which is what makes the frequency-band decomposition of §5.3 well defined on an irregular subset.

Table 1: The two data sources. The simulated dataset is larger by a factor of $\sim$19 in records but covers an order of magnitude fewer species, all of them small. Cross-section extrema are over positive values. P5-P95 denotes the 5th and 95th percentiles of records.

	Experimental (XSC)	**Simulated (line-by-line)**
Molecules / isotopologues	570 / —	53 / 143
Records	1493	28,314
Size strata (mol. / rec.)	199/470 large, 289/780 med., 82/243 small	all small
Temperature (K)	194.3-773 (P5–P95: 278.15–323.15)	9 values, 100-800
Pressure (atm)	0-1	22 values, 10^{-6}-100
Wavenumber grid	native, per record	uniform 0.02 cm^{-1}
$\sigma_{\min}$ / $\sigma_{\max}$ (cm^2/molecule)	1.8×10^{-30} / 7.8×10^{-17}	2.1×10^{-255} / 2.3×10^{-15}

3.4 THE LABEL FLOOR

The two datasets have label floors that differ by 225 decades: experimental spectra have an instrumental detection limit near 10^{-30} cm^2/molecule, while simulated spectra could be down to 10^{-255} due to calculation precision. In log space, this undetectable signal would dominate the loss; training the same regression head on both is therefore a domain mismatch in the label.

We treat the experimental dataset as ground truth and the simulated dataset as an auxiliary source to be aligned to it. Thus, throughout, the training target is $\log_{10}(\sigma + \epsilon)$ with $\epsilon = 10^{-30}$, and experimental points with $\sigma = 0$ are always excluded from the loss. We compare three options for dealing with simulated values below the experimental detection limit in §5.2.

The synthetic data mixture sweep in §5.1 adds a nested fraction (25%, 50%, or 100%) of the simulated dataset to the experimental training folds. Fractions are sampled per record so that chemical coverage is preserved as volume is reduced. The 25% subset is contained in the 50% subset, which is contained in the full set.

4 EXPERIMENTAL PROTOCOL

Every result in §5 is produced under the same protocol: a molecule-wise cross-validation split that forces generalization to unseen chemistry (§4.1), metrics computed per record (§4.2), paired-significance testing at the record level (§4.3), and a one-axis-at-a-time sweep design (§4.4).

4.1 CROSS-VALIDATION

We evaluate on 570 experimental HITRAN molecules under 4-fold molecule-wise cross-validation. Experimental molecules are stratified by atom count and assigned molecule-wise to four folds; fold k is the test set for model k, and the remaining three folds, plus the shared simulation set, form its training set. Grouping is by InChIKey skeleton, falling back to flat canonical SMILES and then to a normalized name, so two catalogue spellings of the same species cannot be split across folds. The assignment is generated once and reused by every configuration.

The split is molecule-wise. A record-wise split would place the same species at 296 K in training and at 323 K in test, and since the dominant variation in a cross section is the molecular identity rather than the thermodynamic state, a model could score well by memorizing spectra and interpolating mildly in (T, P). The global removal of shared species from the simulation (§3.3) serves the same purpose.

4.2 METRICS

We report per-record R^2 and MAE on the $\log_{10}$ target. Every configuration is scored on the same 1493 held-out records, one row per (fold, molecule, isotopologue, T, P), and metrics are computed per record, then pooled across folds. Evaluation is deliberately unweighted even when training is weighted: the training weights encode a signal-versus-baseline prior, and reusing them at evaluation time would put the same thumb on both scales. Points excluded from the loss as unmeasured are likewise excluded from the metric; every point that remains counts equally.

We report median R^2 rather than means because per-record R^2 is unbounded below and its mean is dominated by a handful of pathological records. We therefore report the median R^2 as the headline statistic and treat the mean as an instability diagnostic. All significance testing is rank-based and thus unaffected by the same outliers. Appendix C.3 describes two finer-grained views, an in-band metric and a low/high-frequency band decomposition, and explains why neither displaces the overall numbers.

4.3 PAIRED COMPARISON

Because every configuration is evaluated on an identical record set, comparisons are paired at the record level. Configurations are joined on (fold, molecule, isotopologue, T, P); the join is exact and complete, with no records dropped, so every reported test use $n = 1493$ matched pairs. Within each axis we run a Friedman omnibus test across configurations, followed by pairwise Wilcoxon signed-rank tests with Holm correction over all pairs in that axis.

We report the median paired difference Δ and the matched-pairs rank-biserial correlation r, and interpret results by r, read as small/medium/large at $0.1/0.3/0.5$. Holm-adjusted p-values are tabulated in Appendix C.4. At $n = 1493$, differences far too small to matter reach arbitrarily small p-values, so p establishes only that an effect is real and r establishes whether the effect is large enough to act on. Tests are run independently on R^2 and MAE, and we note any configuration on which the two disagree.

4.4 SWEEP DESIGN AND ITS LIMITS

The design space (Fig. 1 *right*) is traversed one axis at a time, and we do not claim a global optimum. Within an axis, every configuration shares the same training budget, optimizer, schedule, folds, sampled wavenumber points, and every setting on every other axis; only the axis under test varies. Appendix C.2 tabulates the full configuration in force for each axis.

Each axis's conclusion is about the ordering of its own configurations under a fixed surrounding configuration, and absolute numbers are not comparable across axes. Wavenumber encoding and band split were varied jointly, since both control the spectral resolution the decoder can represent and neither is interpretable with the other held arbitrary.

Table 2: Data sources. Each configuration adds a nested fraction of the simulated dataset to the same experimental folds and is tested on the same held-out experimental records. Δ and r are against the experimental-only baseline. R^2/MAE is higher/lower is better, thus corresponding Δ and r are lower/higher are better.

Config.	**Med.** R^2	**Med.** Δ	r	**MAE**	**Med.** Δ	r
Experimental only	0.4634	-	-	0.6164	-	-
25% sim + Exp	0.5875	-0.0878	**-0.865**	0.5534	0.0610	**0.880**
50% sim + Exp	**0.5967**	-0.0940	-0.859	**0.5303**	**0.0645**	0.873
100% sim + Exp	0.5933	**-0.0957**	-0.840	0.5525	0.0643	0.862
Simulated only	-71.5213	72.0976	1.000	9.0574	-8.5192	-1.000

5 RESULTS

Two data-side choices move the metric substantially and four representational and architectural choices do not. Training-set composition is the largest effect on any axis by a wide margin ($r = 0.86$), aligning the simulated label floor to the experimental detection limit is second ($r = 0.39$), and on the remaining four axes, the simplest option wins or ties (Fig. 1 *right*).

5.1 DATA SOURCES

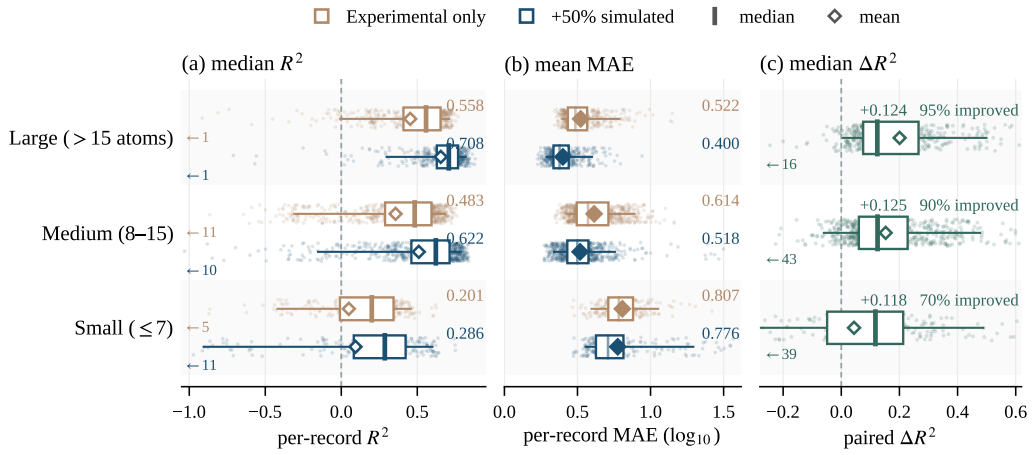


Figure 2: Effect of 50% synthetic augmentation by molecule size. All 53 simulated species have $\leq$ 7 atoms, yet small molecules see only a 3.9% reduction in MAE from the augmentation while large molecules see 23.3%.

Simulated spectra are the strongest lever we found when added to experimental data and a complete failure when used alone. All configurations in Table 2 are tested on the same held-out experimental records.

A model trained only on simulated spectra yields median $R^2 = -71.5$ and MAE = 9.06 in $\log_{10}$ units, meaning predictions are wrong by roughly nine orders of magnitude on a typical point. This is not a degraded version of the right answer; it is a model operating in the wrong label domain, as the label ranges in Table 1 make clear.

The same data, used as an augmentation rather than a substitute, is the largest positive effect in the paper. Adding 50% of the simulated dataset to the experimental folds raises median R^2 from 0.4634 to 0.5967 ($|r| = 0.859$, $p_{\text{Holm}} = 4.44 \times 10^{-181}$) and reduces MAE from 0.6164 to 0.5303 ($r = 0.873$, $p_{\text{Holm}} = 7.13 \times 10^{-187}$). No architectural or representational choice elsewhere in this paper produces an effect of comparable size.

Table 3: Treatment of simulated label values below the experimental detection limit. Experimental non-detections are masked in all three configurations; only the simulated side varies. Δ and r are against the untreated baseline.

Config.	**Med.** R^2	**Med.** Δ	r	**MAE**	**Med.** Δ	r
No treatment	0.6261	-	-	0.5230	-	-
Mask from loss	**0.6442**	**-0.0157**	**-0.390**	**0.5099**	**0.0123**	**0.393**
Clip to floor	0.6246	0.0007	0.007	0.5225	-0.0006	-0.001

The benefit saturates in volume: the 25%, 50% and 100% configurations are statistically close despite a 4-fold range in added records. Because the fractions are nested and sampled per record, all three span the same 53 species, and the benefit saturates once those species are present at all. The obvious reading of that saturation, that the model is simply learning those 53 species, is ruled out by Fig. 2. Every simulated species has ≤ 7 atoms, yet augmentation reduces MAE by 23.3% on large molecules and by only 3.9% on small ones. What the augmentation captures is therefore a prior over the functional form, namely exact, densely sampled, noise-free evidence of how logσ varies with wavenumber and responds to temperature and pressure, which the experimental folds anchor to specific species. This is consistent with the sim-only collapse: the functional prior is correct and the label domain is not, so it is useful only as an auxiliary signal.

We select 50% rather than 100% for stability. At 100%, a median of 0.5933 coexists with a mean of 0.0357, meaning a small number of records are catastrophic.

Fig.6 shows the effect on sample individual records. Under experimental-only training, the prediction tracks the strongest absorption bands but misplaces the baseline level between them. With 50% augmentation, the baseline level is recovered and band contours sharpen, raising R^2 on the illustrated large molecule from 0.5594 to 0.7955 and on the medium molecule from 0.3500 to 0.6436. The small molecule improves from 0.2054 to 0.3210 and remains the weakest of the three.

5.2 Aligning the simulated label floor

Masking simulated points below the experimental detection limit (10^{-30} $\mathrm{cm}^2/\mathrm{molecule}$) from the loss gives a medium improvement over leaving them untouched (r = 0.39 on both metrics), while clipping to the limit is indistinguishable from doing nothing (r = 0.007 on R^2). The result suggests that the benefit comes from removing the unphysical points from the gradient rather than telling the model where the experimental floor sits.

5.3 Wavenumber encoding and band split

A plain positional encoding with a single-head decoder is the best configuration on both metrics, and no Fourier variant improves on it. Encoding and band split (Appendix A.2) are varied jointly (§4.4).

Over-resolving the encoding produces the least stable configuration in the entire sweep. Config. 5 reaches median R^2 0.4832, but mean R^2 -73.1 and MAE 4.87. Pushing the maximum encoded frequency to 242,500, well past what the sampled grid can support gives the decoder a basis in which it can represent structure the data cannot constrain.

Band splitting does not help under any encoding, and hurts most when applied to the best one. Configs. 2, 4 and 6 are at or below their single-head counterparts, and Config. 4 is the clearest loss of the three (r = 0.492 against the axis baseline, and r = 0.641 against Config. 3). The explanation is the same one that made the band decomposition a confirmatory diagnostic in Appendix C.3: at the sampling density of these records, the high band is instrument noise rather than resolvable line structure, so a dedicated head is being asked to predict a component that is not predictable from molecular structure. Giving it its own capacity and its own loss term spends parameters on noise and costs accuracy on the part of the signal that is learnable.

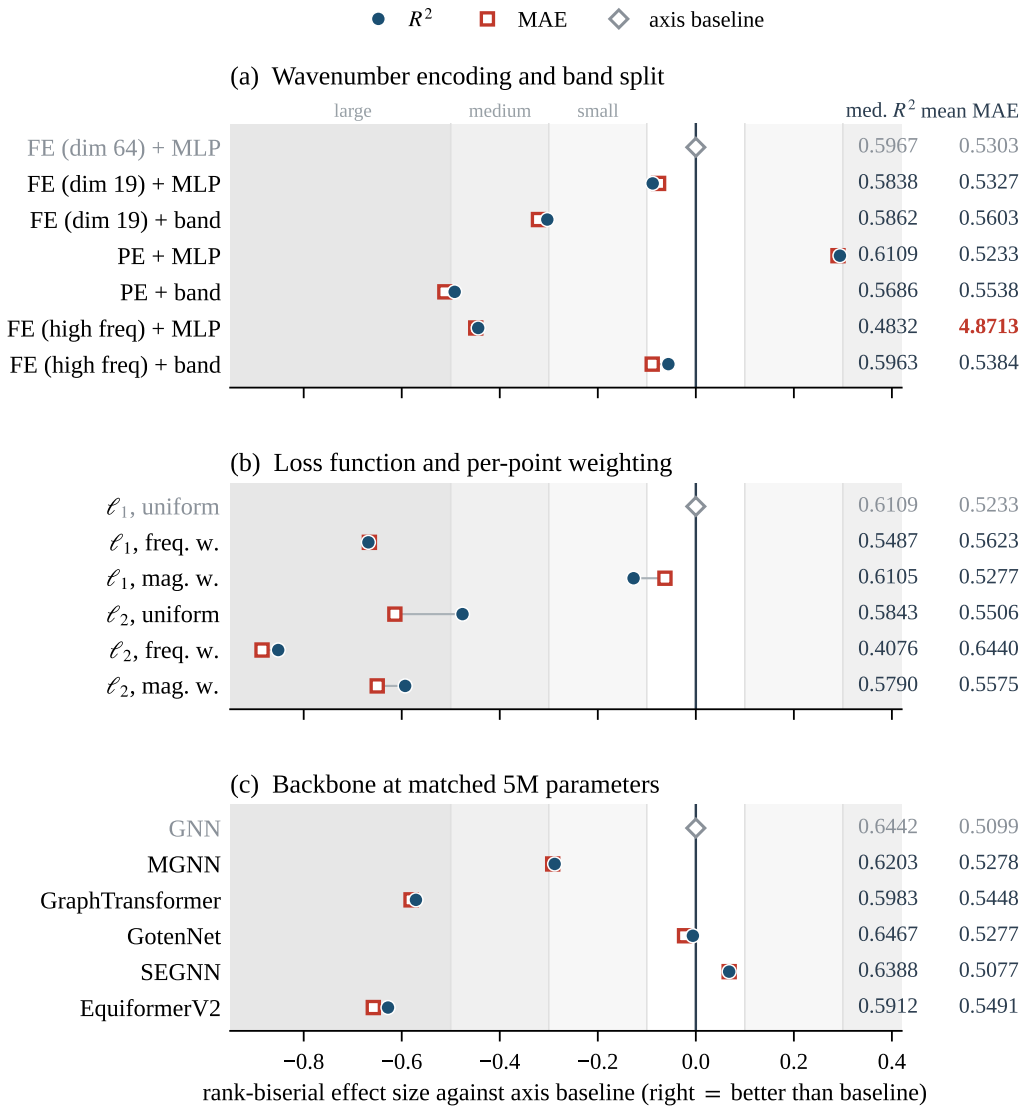


Figure 3: Three design axes, each configuration against its own axis baseline (grey diamond, top row of each panel, $r = 0$). Simpler choice (PE + MLP, ℓ_1 loss with uniform weighting, plain GNN) win or tie with the expressive ones.

5.4 Thermodynamic conditioning

Encoding T and P as Fourier features makes no practical difference compared with normalized scalars, and this is the one axis where the two metrics disagree in sign. Fourier conditioning is marginally ahead on median R^2 (0.6160 against 0.6109, r = 0.076) and slightly behind on MAE (0.5307 against 0.5233, $|r|$ = 0.104). We treat this as a negligible win for the scalar encoding. The absence of a difference is expected, because cross sections vary smoothly in T and P over the ranges present in the experimental archive, with no oscillatory structure for a Fourier basis to capture, and this is also the axis least exercised by the data, since 90% of experimental records sit between 278 K and 323 K.

5.5 Loss and weighting

Plain ℓ_1 with uniform weighting wins, and the ordering on this axis is strict (Fig. 3 b): every departure from it is a significant loss, and the effect sizes grow monotonically with how far the configuration departs. Magnitude weighting is the only near-neighbor, behind by a small effect ($r = 0.127$) that is significant but not practically meaningful.

Both larger effects have the same character, and it is the character described in §5.3. Moving from ℓ_1 to ℓ_2 costs a medium effect on its own (r = 0.476), which is unsurprising for a target with a heavy-tailed residual distribution: squaring the error hands the gradient to the worst-fit records, and those are disproportionately the small molecules the model cannot fit at all (Fig. 2). Frequency weighting (Appendix A.3) costs a large effect on its own (r = 0.668) and compounds with ℓ_2 into the worst configuration in the axis, at median R^2 = 0.4076. It directs capacity toward exactly the high-frequency content that §5.3 found to be unlearnable, and does so by construction, since the statistic it keys on cannot distinguish local line structure from local noise.

5.6 Architecture at a matched 5M-parameter budget

Geometric equivariance and attention give no advantage over a plain GNN on this task. All six backbones are configured to within 5% of 5M parameters by sweeping a width multiplier against the real constructor (Appendix B.3), and share the decoder, training budget, folds, and every other setting.

The two fully equivariant backbones land at opposite ends of the ordering, which is the clearest evidence that geometric equivariance is not what separates these models on this task. SEGNN ties the invariant baseline ($|r|$ = 0.068 on both metrics) while EquiformerV2 is last ($|r|$ = 0.628), so equivariance does not explain the ordering. Attention fares no better: GraphTransformer is fifth of six ($|r|$ = 0.571). GotenNet, GNN, and SEGNN are statistically tied at the top.

Among the three tied backbones, we take GNN as the operating point, because it is tied at the top on both metrics, stable across records, and the cheapest of the three to train. GotenNet illustrates why stability matters. It has the highest median R^2 on the axis (0.6467) but the lowest mean (0.3403); it is right on typical records and occasionally very wrong, which for an emulator intended to be trusted on unseen chemistry is the less desirable of two ways to arrive at a similar average.

We read this null result narrowly, as a statement about our inputs and target rather than about equivariance. Geometries are single RDKit conformers relaxed with MMFF rather than measured structures, so equivariance is being enforced with respect to coordinates that are themselves approximate and, for flexible molecules, arbitrary among many accessible conformers. The target is also a rotationally invariant scalar field indexed by wavenumber, not a tensorial or directional quantity of the kind equivariant message passing was developed for. The result bounds what equivariance buys on this benchmark as constructed; it does not bound what it would buy given measured geometries, conformer ensembles, or directionally structured targets, and we would expect a different answer in each of those settings.

6 Conclusion

We formulated environment-conditioned absorption cross-section prediction as molecule-conditioned continuous regression over wavenumber, and evaluated it on 570 experimental HITRAN species under molecule-wise four-fold cross-validation with record-level paired testing. Across a controlled sweep of data source mixture, label-floor treatment, wavenumber and thermodynamic encoding, loss and weighting, and six backbones at a matched 5M-parameter budget, synthetic data augmentation is the largest contributor: synthetic data alone does not transfer to real data, yet 50% synthetic augmentation lifts median R^2 from 0.46 to 0.60, with the gain accruing mainly to molecules unlike anything in the synthetic dataset. Aligning the simulated label floor to the experimental detection limit by masking rather than clipping is the second-largest lever. Given those two choices, simple encodings win or tie with expressive ones and a plain invariant GNN matches or beats every equivariant and attention-based alternative we tested. We release the folds, the evaluation protocol, and the per-record outputs so that subsequent methods can be compared on the same records under the same paired tests.

AI USE STATEMENT

In this work, we used generative AI tools for polishing writing, finding related work, and debugging. We have not used generative AI tools for research ideation, generating synthetic datasets, or proving mathematical claims.

REFERENCES

Sarp Aykent and Tian Xia. GotenNet: Rethinking Efficient 3D Equivariant Graph Neural Networks. In *The Thirteenth International Conference on LearningRepresentations*, 2025. URL https://openreview.net/forum?id=5wxCQDtbMo.

Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E Smidt, and Boris Kozinsky. E (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature communications*, 13(1): 2453, 2022.

P.F. Bernath. *Spectra of Atoms and Molecules*. Oxford University Press, 2005. ISBN 9780195346459. URL https://books.google.com/books?id=pW3uok6nITMC.

Johannes Brandstetter, Rob Hesselink, Elise van der Pol, Erik J Bekkers, and Max Welling. Geometric and physical quantities improve e (3) equivariant message passing. *arXiv preprint arXiv:2110.02905*, 2021.

Jian Chang and Shuze Zhu. Mgnn: Moment graph neural network for universal molecular potentials. *npj Computational Materials*, 11(1), March 2025. ISSN 2057-3960. doi: 10.1038/s41524-025-01541-5. URL http://dx.doi.org/10.1038/s41524-025-01541-5.

Johannes Gasteiger, Florian Becker, and Stephan Günnemann. Gemnet: Universal directional graph neural networks for molecules. *Advances in Neural Information Processing Systems*, 34:6790–6802, 2021.

I.E. Gordon, L.S. Rothman, R.J. Hargreaves, R. Hashemi, E.V. Karlovets, F.M. Skinner, E.K. Conway, C. Hill, R.V. Kochanov, Y. Tan, P. Wcisło, A.A. Finenko, K. Nelson, P.F. Bernath, M. Birk, V. Boudon, A. Campargue, K.V. Chance, A. Coustenis, B.J. Drouin, J.–M. Flaud, R.R. Gamache, J.T. Hodges, D. Jacquemart, E.J. Mlawer, A.V. Nikitin, V.I. Perevalov, M. Rotger, J. Tennyson, G.C. Toon, H. Tran, V.G. Tyuterev, E.M. Adkins, A. Baker, A. Barbe, E. Canè, A.G. Császár, A. Dudaryonok, O. Egorov, A.J. Fleisher, H. Fleurbaey, A. Foltynowicz, T. Furtenbacher, J.J. Harrison, J.–M. Hartmann, V.–M. Horneman, X. Huang, T. Karman, J. Karns, S. Kassi, I. Kleiner, V. Kofman, F. Kwabia–Tchana, N.N. Lavrentieva, T.J. Lee, D.A. Long, A.A. Lukashevskaya, O.M. Lyulin, V.Yu. Makhnev, W. Matt, S.T. Massie, M. Melosso, S.N. Mikhailenko, D. Mondelain, H.S.P. Müller, O.V. Naumenko, A. Perrin, O.L. Polyansky, E. Raddaoui, P.L. Raston, Z.D. Reed, M. Rey, C. Richard, R. Tóbiás, I. Sadiek, D.W. Schwenke, E. Starikova, K. Sung, F. Tamassia, S.A. Tashkun, J. Vander Auwera, I.A. Vasilenko, A.A. Vigasin, G.L. Villanueva, B. Vispoel, G. Wagner, A. Yachmenev, and S.N. Yurchenko. The hitran2020 molecular spectroscopic database. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 277: 107949, 2022. ISSN 0022-4073. doi: https://doi.org/10.1016/j.jqsrt.2021.107949. URL https://www.sciencedirect.com/science/article/pii/S0022407321004416.

G. Herzberg. *Molecular Spectra and Molecular Structure* -. Number v. 1. Read Books, 2008. ISBN 9781443726085. URL https://books.google.com/books?id=OZoxWcQYijcC.

G. Herzberg and J.W.T. Spinks. *Molecular Spectra and Molecular Structure: Infrared and Raman spectra of polyatomic molecules*. Molecular Spectra and Molecular Structure. Prentice-Hall, 1939. URL https://books.google.com/books?id=vX4fAQAAMAAJ.

Yi-Lun Liao, Brandon Wood, Abhishek Das, and Tess Smidt. Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations, 2024. URL https://arxiv.org/abs/2306.12059.

Charles McGill, Michael Forsuelo, Yanfei Guan, and William H. Green. Predicting infrared spectra with message passing neural networks. *Journal of Chemical Information and Modeling*, 61(6): 2594–2609, June 2021. ISSN 1549-9596. doi: 10.1021/acs.jcim.1c00055. URL https://doi.org/10.1021/acs.jcim.1c00055.

Zhuoran Qiao, Matthew Welborn, Animashree Anandkumar, Frederick R Manby, and Thomas F Miller. Orbnet: Deep learning for quantum chemistry using symmetry-adapted atomic-orbital features. *The Journal of chemical physics*, 153(12), 2020.

L.S. Rothman, I.E. Gordon, R.J. Barber, H. Dothe, R.R. Gamache, A. Goldman, V.I. Perevalov, S.A. Tashkun, and J. Tennyson. Hitemp, the high-temperature molecular spectroscopic database. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 111(15):2139–2150, 2010. ISSN 0022-4073. doi: https://doi.org/10.1016/j.jqsrt.2010.05.001. URL https://www.sciencedirect.com/science/article/pii/S002240731000169X. XVIth Symposium on High Resolution Molecular Spectroscopy (HighRus-2009).

Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. *Advances in neural information processing systems*, 30, 2017.

Kristof Schütt, Oliver Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International conference on machine learning*, pp. 9377–9388. PMLR, 2021.

S. Seager, W. Bains, and J. J. Petkowski. Toward a list of molecules as potential biosignature gases for the search for life on exoplanets and applications to terrestrial biochemistry. *Astrobiology*, 16 (6):465–485, June 2016. doi: 10.1089/ast.2015.1404. URL https://doi.org/10.1089/ast.2015.1404.

Yunsheng Shi, Zhengjie Huang, Shikun Feng, Hui Zhong, Wenjin Wang, and Yu Sun. Masked label prediction: Unified message passing model for semi-supervised classification, 2021. URL https://arxiv.org/abs/2009.03509.

Jonathan Tennyson, Sergei N. Yurchenko, Ahmed F. Al-Refaie, Emma J. Barton, Katy L. Chubb, Phillip A. Coles, S. Diamantopoulou, Maire N. Gorman, Christian Hill, Aden Z. Lam, Lorenzo Lodi, Laura K. McKemmish, Yueqi Na, Alec Owens, Oleg L. Polyansky, Tom Rivlin, Clara Sousa-Silva, Daniel S. Underwood, Andrey Yachmenev, and Emil Zak. The exomol database: Molecular line lists for exoplanet and other hot atmospheres. *Journal of Molecular Spectroscopy*, 327:73–94, 2016. ISSN 0022-2852. doi: https://doi.org/10.1016/j.jms.2016.05.002. URL https://www.sciencedirect.com/science/article/pii/S0022285216300807. New Visions of Spectroscopic Databases, Volume II.

Zihan Zou, Yujin Zhang, Lijun Liang, Mingzhi Wei, Jiancai Leng, Jun Jiang, Yi Luo, and Wei Hu. A deep learning model for predicting selected organic molecular spectra. *Nature Computational Science*, 3(11):957–964, November 2023. ISSN 2662-8457. doi: 10.1038/s43588-023-00550-y. URL https://doi.org/10.1038/s43588-023-00550-y.

A Data Details

A.1 Filtering and grid construction

Simulated thermodynamic grid: the nine temperatures and twenty-two pressures at which line-by-line cross sections were computed.

Temperature (K): 100.0, 200.0, 278.15, 323.15, 400.0, 500.0, 600.0, 700.0, 800.0;

Pressure (bar): 1.0E-5, 2.1544E-5, 4.6416E-5, 1.0E-4, 2.1544E-4, 4.6416E-4, 1.0E-3, 2.1544E-3, 4.6416E-3, 1.0E-2, 2.1544E-2, 4.6416E-2, 1.0E-1, 2.1544E-1, 4.6416E-1, 1.0, 2.1544, 4.6416, 10.0, 21.544, 46.416, 100.0;

The pressures are petitRADTRANS recommended pressures in [bar], and we convert them into [atm].

Molecules excluded from the simulated dataset because they also appear in the experimental archive: Carbon Disulfide (4 isotopologues); Ethane (2 isotopologues); Methyl Fluoride (2 isotopologues); Methyl Iodide; Nitrogen Trifluoride; Nitrous Acid; PFC-14; and Sulfur Hexafluoride.

Exclusion is global and applied before fold assignment.

A.2 Band decomposition

The band split separates a spectrum into a low-frequency component f_L and a high-frequency remainder $f_H = y - f_L$. The decomposition is exact and additive by construction, f_L and f_H sum to the label the model is trained against, so both are comparable to the original target and every metric defined on y applies unchanged to either band.

The cutoff `cutoff_cm` is a period in cm^{-1}: structure broader than the cutoff goes to f_L, narrower to f_H. We use 100 cm^{-1} throughout, which places band contours and envelope in the low band and individual band heads in the high band. A cutoff of 20-30 cm^{-1} would instead make f_L track individual band heads.

The low-pass runs per contiguous measured segment rather than once per record. Each segment is low-passed about its own edges with its own mode count $k \leq 2L_{\text{seg}}/\texttt{cutoff_cm}$. A segment shorter than half the cutoff period supports no mode above DC, so its low band is its mean, which follows from the same formula rather than being special-cased. A full-span simulated record is the one-segment case and behaves as a single transform would. Segments with fewer than $\texttt{min_segment_points} = 8$ valid points likewise receive their mean.

Each segment is mirror-extended, $f(x)$ followed by $f(-x)$, before transforming. The label at 300 cm^{-1} and at 10000 cm^{-1} are not equal, and a plain FFT treats the signal as periodic, so without the even extension the transform sees a step discontinuity at the wrap point and f_L returns with a sawtooth ramp and Gibbs ringing across the full range. The even extension is continuous, which is why a cosine basis is the natural one here. We use a brick-wall mask (`taper` $= 0$, keeping $k \leq k_{\text{cut}}$ and discarding the rest); a raised-cosine edge over $[k_{\text{cut}}/(1+\tau),\ k_{\text{cut}}(1+\tau)]$ is available and removes the sinc ringing that a brick wall induces around strong lines, which is cosmetic for training since f_H cancels it exactly.

Training and evaluation use a random subset of the common grid rather than a uniform sequence, so a transform over the sampled values in storage order would be meaningless. Both available methods work in grid coordinates. The default (`method='fft'`) places the valid samples back on the full uniform grid with linear interpolation across the gaps, applies an rFFT, masks, inverts, and reads back at the sample indices. The alternative (`method='dct'`) performs the same low-pass as a direct projection onto the retained cosine modes, integrated over the sample points, with no full-grid detour and no interpolation; it serves as an independent check that interpolation is not distorting f_L, and the two agree to $\sim$1e-4 in label units at $\texttt{cutoff_cm} = 100$ cm^{-1}.

Split is applied inside the label transform at load time, on the already-transformed label, after the noise floor, simulated-floor treatment, and masking have been applied. The cutoff is therefore switchable per run without rebuilding the dataset, and the two bands always sum to the label the model actually sees. Points outside the valid mask are excluded when building f_L but still re-

ceive an f_L value, so $y - f_L$ remains defined everywhere; this matters because XSC records store $\log_{10}(0) \to -300$ outside the measured window, and admitting that cliff into the transform would swamp every real coefficient.

We use `cutoff_cm` $= 100$ cm^{-1}, `method='fft'`, `taper` $= 0$, `segment_gap_cm` $=$ `cutoff_cm`, `min_segment_points` $= 8$, on the common grid spanning 300-10000 cm^{-1} at 0.02 cm^{-1}. Figure 4 shows the decomposition at three cutoffs for one experimental record.

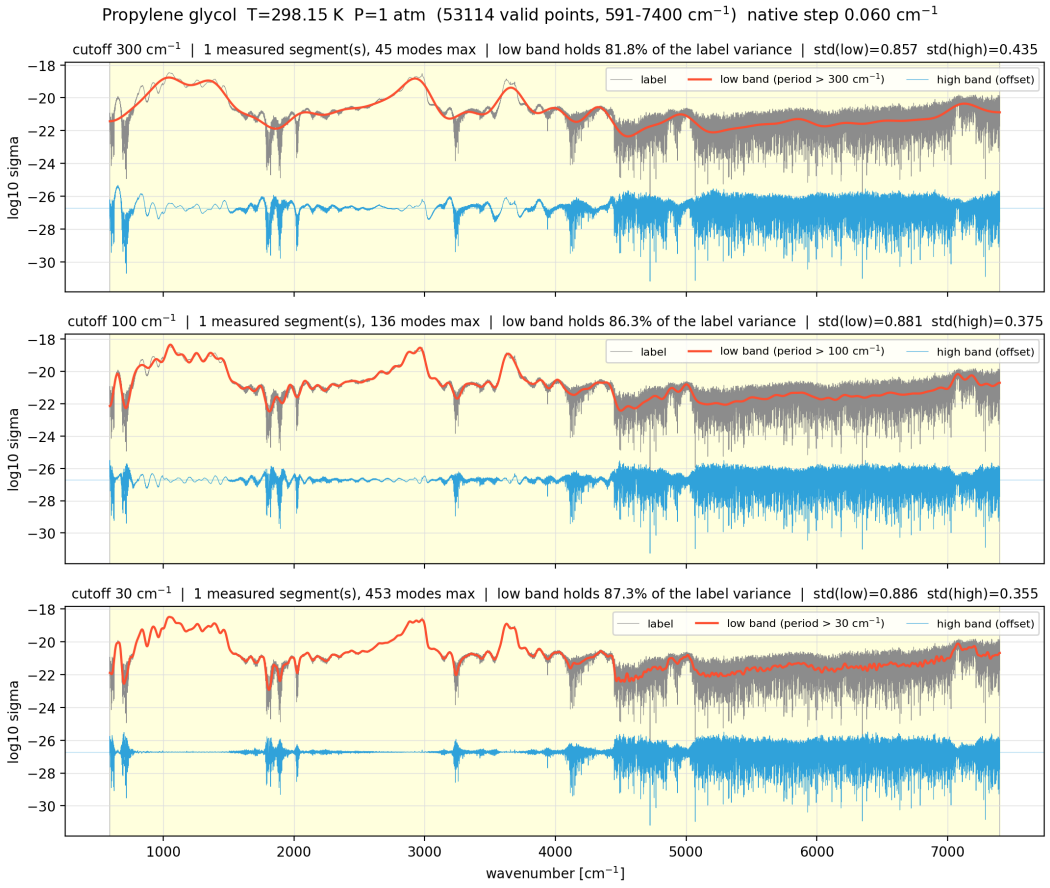


Figure 4: Sample figure for band decomposition at different cutoff.

A.3 Frequency weighting

The frequency weight $w \in [w_{\min}, 1]$ marks, per wavenumber, whether a real spectral feature is present, and is used by the frequency-weighted loss configurations of §5.5 and to define the in-band point set of Appendix C.3. It is computed on the linear cross section, not the $\log_{10}$ target.

A baseline is estimated by grey-scale morphological opening with a structuring element of `window_cm` $= 400$ cm^{-1}, which is peak-robust and costs $O(n)$ regardless of window size. The residual above that baseline is clipped at zero, and a single global median absolute deviation over the record supplies the noise scale. Writing r for the clipped residual, $\tilde{r}$ for its median and $\mathrm{MAD}(r)$ for its median absolute deviation,

$$w_{\text{raw}} = \mathrm{clip}\left(\frac{r - \tilde{r}}{k_{\text{MAD}}\,\mathrm{MAD}(r)},\ 0,\ 1 \right), \tag{2}$$

with $k_{\text{MAD}} = 6$. The raw weight is then smoothed with a uniform filter of width `smooth_cm` $= 20$ cm^{-1}, giving soft edges around each feature, and clipped from below at `min_weight` $= 0.05$ so that no point is ever fully excluded.

The same core is used for both, and its behavior differs appropriately with the data. For experimental records the baseline away from a band is small but noisy, the global MAD sets the scale at which a fluctuation counts as a feature, and the weight is doing genuine signal-versus-noise discrimination. For simulated records there is no measurement noise: the Voigt line shape is truncated at a fixed distance from each line center, so the coefficient is exactly zero away from every line. The MAD then collapses toward the numerical floor, any nonzero point saturates to $w \approx 1$, and the weight degenerates to a smoothed line-present mask.

For simulated data, every (T, P) spectrum for a given molecule and isotopologue is computed on an identical wavenumber grid, and line positions barely move with temperature or pressure, only strengths and widths do, since temperature changes Boltzmann populations and pressure changes collisional broadening. The weight is therefore computed once per isotopologue folder, on the pointwise maximum across all its (T, P) files, and reused for every condition. Beyond being roughly two orders of magnitude cheaper than per-file computation, this is the more defensible choice physically: a line that is weak at 100 K and strong at 500 K should count as a real feature at every temperature, not only at those where it happens to clear a per-file threshold.

We use `window_cm` $= 400$ cm^{-1}, $k_{\text{MAD}} = 6$, `smooth_cm` $= 20$ cm^{-1}, `min_weight` $= 0.05$, `mode='per_molecule'` for the simulated dataset. Figure 5 shows the resulting weight against a measured cross section.

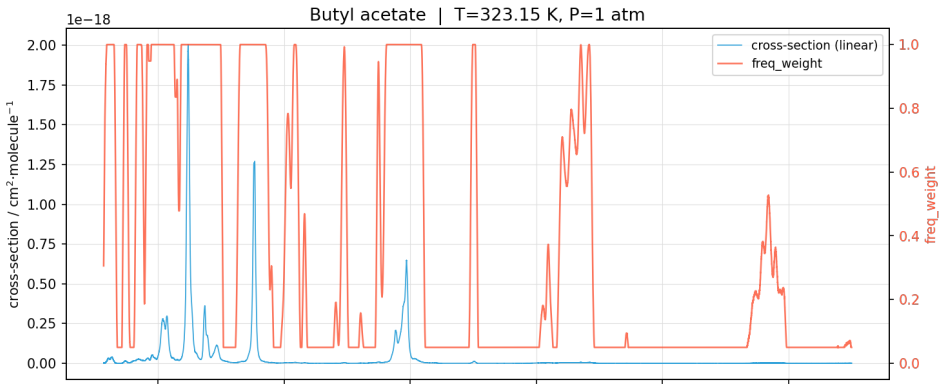


Figure 5: Sample frequency weight against a measured cross section for Butyl acetate.

B Model and Training Details

B.1 Shared decoder

Every backbone is paired with the same decoder, so the architecture axis varies the encoder alone. The encoder consumes structure only and returns a molecule embedding of dimension 128; wavenumber, temperature and pressure enter at decode time as coordinates. This is the architectural expression of the task formulation in §3: because conditions are query coordinates rather than node features, one trained model can be evaluated at any (ν, T, P) without re-encoding the molecule, and the molecule embedding is computed once per record rather than once per point.

The decoder concatenates the molecule embedding, the encoded wavenumber, and the encoded (T, P) into a single vector per query point, and maps it through a four-layer MLP with SiLU activations and widths $(d_{\text{in}} \to h \to h \to h/2 \to 1)$, with $h =$ `decoder_hidden_dim` $= 256$. Wavenumbers are normalized to $[0, 1]$ over 300-10000 cm^{-1} before encoding.

Two encoders are available and are compared in §5.3. Positional is the NeRF dyadic encoding $\gamma(x) = [x, \sin(2^k x), \cos(2^k x)]$ for $k = 0 \dots L-1$, with 2^k as angular frequencies and no factor of 2π. Fourier uses L log-spaced frequencies between `wavenumber_min_freq` and `wavenumber_max_freq`. The two have identical output dimension for the same L and are drop-in swappable, but they differ in what the comparison actually tests: for the positional encoding, bandwidth is tied to L, so the top band 2^{L-1} corresponds to $2^{L-1}/2\pi$ cycles across the span, while

for the Fourier encoding bandwidth is set independently by `max_freq`. The axis therefore compares bandwidth and spacing jointly rather than spacing alone.

At $L = 19$ the positional encoding reaches $2^{18}/2\pi \approx 41721$ cycles across the span, a finest period of 0.23 cm^{-1}, which is matched by a Fourier encoding at `max_freq` $= 43000$. And beyond $L = 20$ the positional encoding is numerically meaningless: in float32, $\sin(2^{L-1}x)$ for $x \in [0, 1]$ carries a maximum error of 1.6e-2 at $L = 20$, rising to the function's full range of 2.0 by $L = 22$. The implementation refuses $L > 20$, which is why the over-resolved configurations use the Fourier encoder.

Temperature is normalized linearly over 100-800 K and pressure logarithmically over $\log_{10} P \in [-6, 2.5]$, both clamped to $[0, 1]$. With `tp_embed_dim` $= 0$ the decoder receives the two normalized scalars directly; with `tp_embed_dim` > 0 each is expanded by a Fourier encoding with `freq_max` $= 16$, far below the wavenumber encoder's bandwidth because T and P carry no fine structure the encoding exists to let the decoder represent a non-monotonic response such as line broadening.

The two-head variant replaces the single output with separate low- and high-band heads whose predictions sum to the spectrum, and bounds the low head's wavenumber bandwidth to what the 100 cm^{-1} cutoff admits, so the low head cannot represent structure that belongs to the high band by construction. Both heads are trained against their respective components of the decomposed label (Appendix A.2).

B.2 Backbones

All six encoders are reimplemented against a common interface: a structure-only encoder returning a molecule embedding. Atom features are 23-dimensional throughout, 22 electronic-configuration features indexed by atomic number, plus the isotopologue index, and bond type and bond order are embedded and supplied as edge features. Models that build their own radius graph have the stored covalent bond types reconciled onto it, with radius-graph edges absent from the bond list marked as long-range. Deviations from each published architecture are listed below; they follow from this shared interface and the matched parameter budget, so §5.6 compares these implementations rather than the original models as published.

- **GNN.** Invariant message passing. Each layer concatenates the two endpoint features with the edge feature, passes them through a two-layer MLP to form the message, and updates the node with a residual MLP over the concatenation of its prior state and the aggregated message; mean pooling gives the molecule embedding.
- **MGNN.** Rotation-invariant rather than equivariant: angular information enters only as scalar moments, dot products of unit bond vectors sharing a center, $\cos^n \theta$ for $n = 1, 2$, so no steerable $\ell > 0$ tensor is ever carried between layers. The radial basis is Chebyshev polynomials with a smooth cutoff envelope. Deviations: node embedding is a linear map over the 23-dimensional atom features rather than the paper's atomic-number embedding; mean pooling rather than sum pooling, matching the other encoders, since sum pooling is specific to extensive scalars such as total energy; and the paper has no thermodynamic conditioning at all.
- **GraphTransformer.** Multi-head dot-product attention over the graph, with edge features entering keys and values and a gated residual per node. Attention is local, restricted to graph edges, with no global attention and no structural or spatial positional encodings, so this configuration tests attention-weighted local message passing rather than the molecular graph-transformer lineage that adds spatial encodings over all node pairs.
- **GotenNet.** Semi-equivariant. Deviation: fully connected tensor products, whose weight count grows as the product of the three multiplicities per path, are replaced by a depthwise tensor product with one weight per channel followed by a cheap linear mix. This is the parameter-efficiency mechanism the original work relies on, and it is what makes the target budget reachable.
- **SEGNN.** Steerable $E(3)$-equivariant message passing over irreducible representations, with steerable node and edge attributes from spherical harmonics of the relative positions. Hidden irreps are allocated by a balanced-irreps scheme at the given $\ell_{\max}$, and normalization is instance-wise over irrep blocks.

Table 4: Backbone configurations at the matched 5M-parameter budget

Backbone	`hidden`	Total params.
GNN	486	4,992,027
MGNN	136	4,963,585
GraphTransformer	432	4,905,321
GotenNet	189	4,980,828
SEGNN	512	4,974,003
EquiformerV2	368	4,923,945

Table 5: Training hyperparameters, held fixed across every configuration of every axis.

Setting	Value
Optimizer	Adam
Learning rate	1e-3
Schedule	MultiStepLR, $\gamma = 0.1$ at $E/5$, $E/2$, $E/1.1$
Epochs E	150
Batch size	64
Points per record	60000 sampled
Seed	42
Molecule embedding dim	128
Decoder hidden dim	256
Encoder layers	3
Radius cutoff	5 Ångström
Loss (baseline)	ℓ_1, uniform weighting
Evaluation cadence	every 10 epochs

- **EquiformerV2.** Equivariant attention with SO(2) linear layers in place of full tensor products, degree-wise feedforward with radial gating, separate $\ell = 0$ and $\ell > 0$ paths, and per-block learnable layer scale. Edge embeddings combine a radial basis with an envelope and spherical harmonics up to $\ell_{\max}$, concatenated with the bond-type and bond-order embeddings before attention.

B.3 Parameter matching

Backbones expose different width controls, `hidden_dim` for the invariant and attention models, `hidden_features` together with $\ell_{\max}$ for the steerable ones, and parameter count is not a smooth function of any of them, since irrep multiplicities quantize. We therefore sweep the width control against the real constructor, instantiating the full model and counting parameters at each step, and select the setting whose realized total falls closest to the target within a tolerance of 5%. The decoder is held fixed across all backbones, so the sweep adjusts encoder capacity only; realized totals are reported in Table 4.

B.4 Training hyperparameters

Every configuration of every axis uses the settings in Table 5. The learning-rate schedule is a multi-step decay with $\gamma = 0.1$ at milestones placed at $E/5$, $E/2$ and $E/1.1$ epochs, so the final $\sim$10% of training runs at 10^{-3} of the initial rate.

Table 6: Configuration in force for each axis. The axis varied within a axis is marked; all other settings are held fixed across that axis' configurations.

axis	Data sources	Floor	Wavenum. enc.	Band	T/P enc.	Loss	Backbone
1	**varied**	no treat	FE (dim 64)	MLP	scalar	ℓ_1, uni.	GNN
2	+50% sim	**varied**	PE	MLP	scalar	ℓ_1, uni.	GNN
3	+50% sim	no treat	**varied**	**varied**	scalar	ℓ_1, uni.	GNN
4	+50% sim	no treat	PE	MLP	**varied**	ℓ_1, uni.	GNN
5	+50% sim	no treat	PE	MLP	scalar	**varied**	GNN
6	+50% sim	mask	PE	MLP	scalar	ℓ_1, uni.	**varied**

C Full Results

C.1 Sample prediction result

Fig. 6 shows sample prediction result of the same large, medium, and small molecules at the same thermodynamic condition for experimental-only configuration and with 50% synthetic data augmentation.

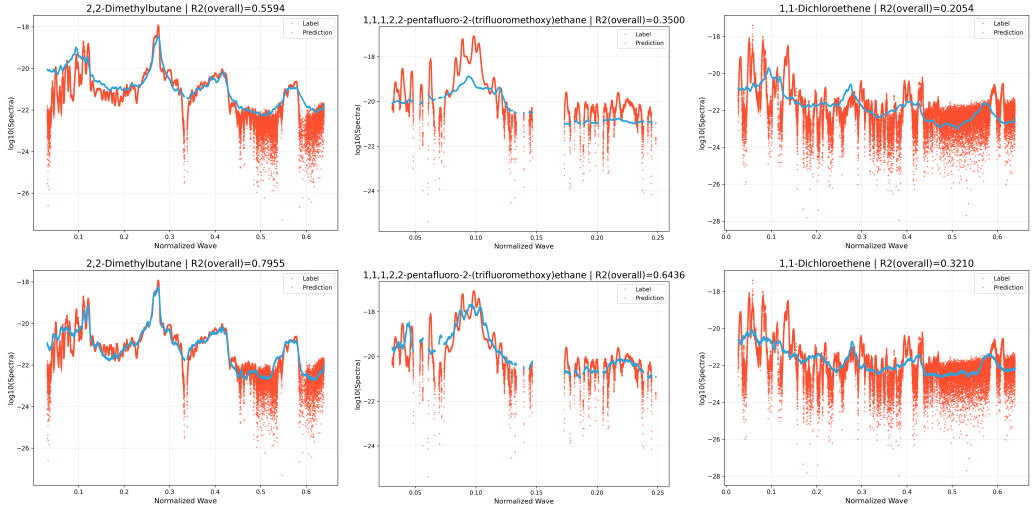


Figure 6: Sample predictions of the same large (*left*), medium (*middle*), and small (*right*) molecules at the same thermodynamic condition for experimenatl configuration (*top*) and with 50% synthetic data augmentation (*bottom*).

C.2 Configuration in force for each axis

As described in §4, the axis do not form a single cumulative trajectory. Within each axis every configuration shares all settings except the axis under test; across axis the surrounding configuration differs, for the reasons given there. Table 6 records the full configuration in force for each axis so that no absolute number in the main text is read against a baseline it was not measured under.

C.3 Supplementary Metric

We considered restricts scoring to points inside a real spectral feature (in-band), gated by the per-point frequency weight. Its two forms disagree systematically. Removing baseline points shrinks the variance that R^2 normalizes against, so on records the model already fits well the in-band R^2 rises slightly, while on poorly fit records it falls sharply; When pooled, the tail dominates and in-band

Table 7: axis 1, mean R^2, mean rank, and best-configuration percentage

	Mean R^2	Mean Rank (R^2)	Best % (R^2)	Mean Rank (MAE)	Best % (MAE)
0 Experimental only	0.3376	3.587	4.2%	3.590	4.7%
1 25% sim + Exp	0.3451	2.206	30.3%	2.190	29.7%
2 50% sim + Exp	**0.4776**	**2.082**	28.2%	**2.100**	**37.4%**
3 100% sim + Exp	0.0357	2.126	**37.2%**	2.121	28.2%
4 Simulated only	-81.3166	4.999	0.0%	4.999	0.0%

Table 8: axis 1, pairwise Wilcoxon ranking (Holm-corrected over 10 pairs)

A vs B	p_{Holm} (R^2)	r (R^2)	p_{Holm} (MAE)	r (MAE)
0 vs 1	1.06e-183	-0.865	5.04e-190	0.880
0 vs 2	4.44e-181	-0.859	7.13e-187	0.873
0 vs 3	3.52e-173	-0.840	3.16e-182	0.862
0 vs 4	1.40e-244	1.000	1.40e-244	-1.000
1 vs 2	5.27e-04	-0.112	1.32e-03	0.105
1 vs 3	8.02e-02	-0.061	1.66e-02	0.079
1 vs 4	1.40e-244	1.000	1.40e-244	-1.000
2 vs 3	1.19e-01	0.047	7.97e-01	-0.008
2 vs 4	1.40e-244	1.000	1.40e-244	-1.000
3 vs 4	3.53e-242	0.995	1.40e-244	-1.000

R^2 looks worse than overall even as in-band MAE improves. A view whose two metrics move in opposite directions is a poor arbiter between configurations, so we do not use it as one.

We then considered splits both label and prediction with an identical low-pass at a 100 cm^{-1} period cut, applied per contiguous measured segment. On these records the low band is the spectrum, band contours and envelope carrying roughly 91% of the label variance, and the high band is dominated by instrument noise rather than resolved line structure, which the native sampling of most XSC records cannot support in the first place. The measurements bear this out: R^2(high) sits near zero for every configuration we trained, indicating that the residual high-frequency content is not predictable from molecular structure, while R^2(low) tracks R^2(overall) closely, configuration by configuration, as does the corresponding pair of MAEs. The decomposition therefore confirms the overall metric rather than correcting it, the overall number is not concealing a model that fits an envelope and misses the chemistry, because at this resolution there is no separable line structure in the label to miss.

C.4 Complete sweep tables

Main-text tables report median R^2, MAE, and corresponding effect size against the axis baseline. The tables below add mean R^2, mean rank, best-configuration percentage, and pairwise Wilcoxon ranking with Holm-adjusted p-values, for both metrics and for every configuration.

Table 9: axis 2, mean R^2, mean rank, and best-configuration percentage

	Mean R^2	Mean Rank (R^2)	Best % (R^2)	Mean Rank (MAE)	Best % (MAE)
0 No treatment	0.4877	2.161	24.8%	2.153	26.3%
1 Mask from loss	**0.5020**	**1.658**	**53.2%**	**1.666**	**51.8%**
2 Impute at floor	0.4858	2.181	22.0%	2.182	21.9%

Table 10: axis 2, pairwise Wilcoxon ranking (Holm-corrected over 3 pairs)

A vs B	p_{Holm} (R^2)	r (R^2)	p_{Holm} (MAE)	r (MAE)
0 vs 1	1.72e-38	-0.390	3.65e-39	0.393
0 vs 2	8.25e-01	0.007	9.78e-01	-0.001
1 vs 2	6.95e-37	0.381	1.97e-41	-0.405

Table 11: axis 3, mean R^2, mean rank, and best-configuration percentage

	Mean R^2	Mean Rank (R^2)	Best % (R^2)	Mean Rank (MAE)	Best % (MAE)
0 FE (dim 64) + MLP	0.4776	3.585	14.5%	3.563	15.7%
1 FE (dim 19) + MLP	0.4741	3.669	21.6%	3.612	23.2%
2 FE (dim 19) + band	0.3887	4.386	7.0%	4.420	7.4%
3 PE + MLP	**0.4778**	**2.957**	**24.7%**	**2.912**	**25.1%**
4 PE + band	0.4514	5.066	5.0%	5.070	4.4%
5 FE (high freq) + MLP	-73.0844	4.514	15.5%	4.526	14.8%
6 FE (high freq) + band	0.4562	3.821	11.6%	3.897	9.4%

Table 12: axis 3, pairwise Wilcoxon ranking (Holm-corrected over 21 pairs)

A vs B	p_{Holm} (R^2)	r (R^2)	p_{Holm} (MAE)	r (MAE)
0 vs 1	9.93e-03	0.088	1.05e-02	-0.076
0 vs 2	3.28e-23	0.303	6.12e-26	-0.321
0 vs 3	6.64e-22	-0.294	2.39e-21	0.290
0 vs 4	1.04e-59	0.492	1.57e-64	-0.512
0 vs 5	7.64e-49	0.444	5.80e-50	-0.449
0 vs 6	6.22e-02	0.056	5.58e-03	-0.089
1 vs 2	2.53e-10	0.196	2.33e-13	-0.226
1 vs 3	2.86e-12	-0.216	1.99e-12	0.217
1 vs 4	2.25e-29	0.342	2.17e-36	-0.382
1 vs 5	1.81e-29	0.343	1.26e-31	-0.356
1 vs 6	4.02e-02	0.069	1.33e-04	-0.122
2 vs 3	2.97e-70	-0.534	2.25e-78	0.565
2 vs 4	8.05e-33	0.363	2.93e-35	-0.376
2 vs 5	6.14e-27	0.327	5.40e-24	-0.308
2 vs 6	7.14e-16	-0.248	9.27e-18	0.263
3 vs 4	1.27e-100	0.641	6.44e-111	-0.673
3 vs 5	2.63e-82	0.579	7.13e-88	-0.598
3 vs 6	1.90e-38	0.393	2.18e-50	-0.452
4 vs 5	3.15e-10	0.194	3.13e-09	-0.184
4 vs 6	4.47e-73	-0.545	5.72e-73	0.545
5 vs 6	6.12e-43	-0.416	1.98e-39	0.398

Table 13: axis 4, mean R^2, mean rank, and best-configuration percentage

	Mean R^2	Mean Rank (R^2)	Best % (R^2)	Mean Rank (MAE)	Best % (MAE)
0 Scalar (dim 0)	**0.4778**	**1.460**	**54.0%**	**1.459**	**54.1%**
1 Fourier (dim 4)	0.4612	1.540	46.0%	1.541	45.9%

Table 14: axis 4, pairwise Wilcoxon ranking (Holm-corrected over 1 pairs)

A vs B	p_{Holm} (R^2)	r (R^2)	p_{Holm} (MAE)	r (MAE)
0 vs 1	1.12e-02	0.076	5.09e-04	-0.104

Table 15: axis 5, mean R^2, mean rank, and best-configuration percentage

	Mean R^2	Mean Rank (R^2)	Best % (R^2)	Mean Rank (MAE)	Best % (MAE)
0 ℓ_1, uniform	**0.4778**	**2.321**	**35.5%**	**2.254**	36.1%
1 ℓ_1, freq w.	0.4348	3.999	8.6%	3.783	9.2%
2 ℓ_1, mag w.	0.4749	2.567	31.9%	2.380	**36.4%**
3 ℓ_2, uniform	0.4638	3.094	13.9%	3.427	9.6%
4 ℓ_2, freq w.	0.3210	5.353	2.9%	5.353	6.2%
5 ℓ_2, mag w.	0.4498	3.666	7.3%	3.804	6.2%

Table 16: axis 5, pairwise Wilcoxon ranking (Holm-corrected over 15 pairs)

A vs B	p_{Holm} (R^2)	r (R^2)	p_{Holm} (MAE)	r (MAE)
0 vs 1	1.14e-109	0.668	3.49e-109	-0.666
0 vs 2	2.05e-05	0.127	7.19e-02	-0.063
0 vs 3	1.90e-56	0.476	4.22e-93	-0.614
0 vs 4	1.61e-177	0.852	1.17e-191	-0.885
0 vs 5	1.29e-86	0.593	5.87e-104	-0.650
1 vs 2	9.84e-89	-0.600	1.13e-91	0.609
1 vs 3	9.45e-44	-0.418	2.18e-10	0.195
1 vs 4	2.04e-138	0.751	8.78e-160	-0.808
1 vs 5	6.80e-12	-0.208	3.75e-01	0.027
2 vs 3	3.31e-21	0.286	1.16e-60	-0.494
2 vs 4	1.20e-180	0.859	2.18e-195	-0.894
2 vs 5	4.65e-75	0.551	4.45e-110	-0.669
3 vs 4	2.30e-161	0.812	1.63e-164	-0.820
3 vs 5	1.85e-25	0.315	1.43e-15	-0.244
4 vs 5	3.30e-149	-0.780	1.24e-151	0.786

Table 17: axis 6, mean R^2, mean rank, and best-configuration percentage

	Mean R^2	Mean Rank (R^2)	Best % (R^2)	Mean Rank (MAE)	Best % (MAE)
0 GNN	0.5020	2.854	24.1%	2.852	24.3%
1 MGNN	0.4835	3.589	13.1%	3.575	13.9%
2 GraphTransformer	0.4377	4.449	4.4%	4.384	4.8%
3 GotenNet	0.3403	2.937	23.9%	2.977	22.0%
4 SEGNN	**0.5035**	**2.694**	**28.6%**	**2.688**	**28.8%**
5 EquiformerV2	0.4296	4.478	5.9%	4.524	6.2%

Table 18: axis 6, pairwise Wilcoxon ranking (Holm-corrected over 15 pairs)

A vs B	p_{Holm} (R^2)	r (R^2)	p_{Holm} (MAE)	r (MAE)
0 vs 1	2.89e-21	0.288	9.24e-22	-0.292
0 vs 2	2.41e-80	0.571	4.37e-83	-0.581
0 vs 3	8.35e-01	0.006	4.41e-01	-0.023
0 vs 4	6.86e-02	-0.068	4.75e-02	0.068
0 vs 5	7.57e-97	0.628	3.20e-106	-0.658
1 vs 2	1.99e-30	0.348	2.61e-27	-0.329
1 vs 3	4.12e-18	-0.265	1.65e-17	0.260
1 vs 4	7.94e-36	-0.379	3.60e-39	0.396
1 vs 5	4.59e-35	0.374	3.64e-36	-0.380
2 vs 3	2.12e-63	-0.506	1.45e-67	0.523
2 vs 4	2.29e-85	-0.589	1.60e-93	0.617
2 vs 5	1.42e-01	0.054	1.63e-02	-0.083
3 vs 4	9.80e-03	-0.091	2.70e-04	0.119
3 vs 5	2.93e-59	0.489	4.43e-70	-0.533
4 vs 5	9.44e-83	0.580	5.44e-92	-0.612