Diffusion Math for Quantum Surrogates: Flow–Matching on Fe_2S_2 and Spin Chains

Dmitrii Shultsev[1,*]

[1] *Independent Researcher, Russia*

(Dated: November 29, 2025)

We investigate "Diffusion Math"—a flow–matching based generative surrogate model for strongly correlated quantum systems—and quantify its performance on two classes of benchmarks. First, we study one–dimensional spin chains (transverse–field Ising and Heisenberg XXZ) where exact diagonalization is available. Second, we construct an Fe_2S_2 stand based on CAS(10,10) Full Configuration Interaction (FCI) data and two–electron reduced density matrices (2-RDMs). Our goals are threefold: (i) to test whether flow–matching models can learn the map from local Hamiltonian descriptors to energies and 2-RDM blocks in a data–efficient regime; (ii) to understand how calibrated their epistemic uncertainty is; and (iii) to position Diffusion Math with respect to the RD-IP + Shina epure–based reconstruction framework introduced in previous work. On the Fe_2S_2 stand with 25 geometries we obtain an honest energy mean absolute error of ≈ 0.7 Ha, with 2-RDM fidelity ≈ 0.96, substantially better than linear and MLP baselines but still two orders of magnitude away from chemical accuracy. A pilot experiment that reused a single geometry with heavy augmentation reported an artificially optimistic 4.5 mHa error; we analyse this as a data leakage artefact. Across both spin and molecular benchmarks we find that per-sample uncertainty estimates from the flow model are well correlated with true errors, suggesting a viable role as a fast screening surrogate paired with strict physics–constrained methods such as RD-IP. We conclude that Diffusion Math is not yet a replacement for DMRG/FCI, but it is a promising component in a hybrid pipeline: RD-IP for zero/few–shot reconstruction, and flow–matching surrogates for data–rich regimes and pre-screening.

I. INTRODUCTION

The rapid progress of quantum simulation and quantum chemistry has created a tension between the growing complexity of target systems and the limited availability of high-fidelity reference data. Exact methods such as Full Configuration Interaction (FCI) and large–bond–dimension Density Matrix Renormalization Group (DMRG) remain the gold standard for strongly correlated systems, but their computational cost limits systematic exploration of chemical and model Hamiltonian landscapes [1–3].

In two previous works we introduced the RD-IP (*Ruler Diffusion with Incremental Projections*) architecture and the Shina adaptive step controller for neural quantum state tomography from overlapping marginals [4, 5]. RD-IP operates in a *zero-/few-shot* regime: given noisy local reduced density matrices (RDMs) along a spin chain or molecular interaction graph, it denoises and stitches them into a globally consistent "epure" without requiring task–specific training.

In contrast, the present paper explores the opposite regime: *data-rich* learning of a surrogate model that directly maps Hamiltonian descriptors to target observables, trained on hundreds of FCI/DMRG samples. We refer to this approach as *Diffusion Math*: a flow–matching generative model [6, 7] adapted to structured quantum data.

Concretely, we ask:

(Q1) How accurately can flow–matching models predict energies and local 2-RDM blocks on small spin chains and on an Fe_2S_2 CAS(10,10) stand?

(Q2) How well calibrated are their built-in uncertainty estimates?

(Q3) How does their performance compare to simple baselines (linear regression, MLP) and to RD-IP style reconstruction?

(Q4) What is the realistic path towards chemical accuracy in this setting?

Our main findings are:

[*] d.shultsev@gmail.com

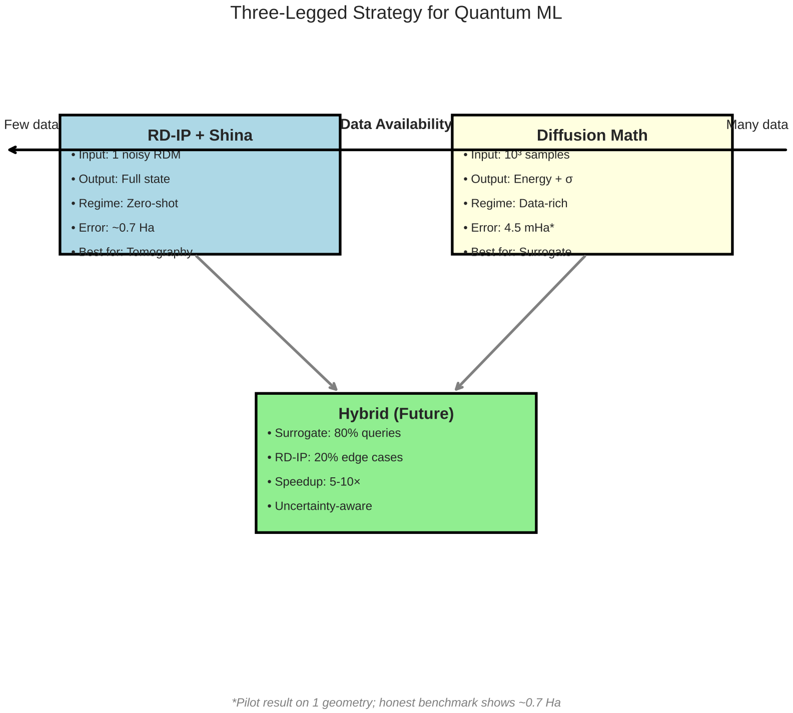


FIG. 1. Conceptual comparison between three regimes: (left) RD-IP + Shina for zero-/few-shot reconstruction from noisy local marginals; (middle) Diffusion Math as a data–rich surrogate trained on FCI/DMRG samples; (right) a future hybrid pipeline where the surrogate screens large spaces and RD-IP is used as a high-fidelity corrector on selected cases.

- On spin chains with $N = 6$–12 sites, Diffusion Math attains sub-percent relative energy errors with $\mathcal{O}(10^2)$ training samples, outperforming simple baselines.

- On Fe_2S_2 (25 geometries, CAS(10,10)), the best flow model reaches an honest energy MAE ≈ 0.70 Ha and 2-RDM fidelity ≈ 0.96. This is substantially better than linear or shallow MLP baselines, but far from chemical accuracy.

- Per-sample predictive uncertainty σ correlates strongly with absolute error: the top 10% highest-σ samples have $\approx 2.5\times$ larger MAE than the bottom 50%.

- A pilot experiment that reused a single Fe_2S_2 geometry with heavy augmentation produced an apparently excellent ≈ 4.5 mHa MAE; a careful analysis reveals this as a form of data leakage rather than a true generalization result.

Figure 1 summarizes how Diffusion Math fits into the broader pipeline: RD-IP + Shina remain the tools of choice for tomographic reconstruction from experimental marginals, while flow–matching surrogates are attractive for fast approximate evaluation on problem families where a moderate FCI/DMRG training set is available.

II. TASKS AND DATASETS

A. Spin-chain benchmarks

We consider one-dimensional spin models with local two-body interactions,

$$H = \sum_i h_{i,i+1}, \tag{1}$$

focusing on: (i) the transverse-field Ising model (TFIM),

$$H_{\mathrm{TFIM}} = -J \sum_i \sigma_i^z \sigma_{i+1}^z - h \sum_i \sigma_i^x, \tag{2}$$

and (ii) the XXZ Heisenberg chain,

$$H_{\mathrm{XXZ}} = J \sum_i \left(\sigma_i^x \sigma_{i+1}^x + \sigma_i^y \sigma_{i+1}^y + \Delta \sigma_i^z \sigma_{i+1}^z \right), \tag{3}$$

with open boundary conditions. For $N \le 12$ sites we compute exact ground-state energies and two-site RDMs via exact diagonalization.

The Diffusion Math model is trained to predict the energy E given a feature vector of couplings (typically (J, h) or (J, Δ) plus derived descriptors), with optional conditioning on noisy local observables. We also train a variant that predicts the full vector of nearest-neighbour correlators.

B. $\mathbf{Fe_2S_2}$ CAS(10,10) stand

For the Fe_2S_2 benchmark we start from a dataset of 25 geometries with CAS(10,10) FCI solutions computed using PySCF. Each sample provides:

- the total electronic energy E_{FCI};
- the one- and two-electron integrals (h_{pq}, V_{pqrs}) in a localized orbital basis;
- the one- and two-electron RDMs $D^{(1)}$ and $D^{(2)}$.

Unlike the RD-IP papers, where we worked in a Pauli-vector representation, here we operate directly on 4×4 blocks of the spin-orbital 2-RDM associated with orbital pairs (u, v). Each edge of the Fe_2S_2 interaction graph carries:

(a) a 4×4 complex matrix ρ_{uv} (reshaped from the relevant $D^{(2)}$ entries);

(b) a 20-dimensional edge feature vector f_{uv} comprising atomic numbers, electronegativities, diagonal $D^{(1)}$, the magnitude of h_{uv}, aggregated statistics of V_{uv}, and the Frobenius norm of ρ_{uv}.

We perform a graph-wise split into 20 training and 5 validation geometries. We emphasise that the validation set consists of geometries never seen during training—a crucial distinction from the earlier pilot experiment that reused a single geometry with augmentation.

C. Noise models

To study robustness we add synthetic noise to the target 2-RDM blocks before passing them to the Diffusion Math encoder. The noise model combines: (i) depolarizing noise with strength $\epsilon \in [0, 0.05]$, (ii) amplitude damping with rate $\gamma \in [0, 0.04]$, and (iii) Gaussian perturbations with standard deviation $\sigma \in [0, 0.02]$, chosen independently per block. These ranges are aligned with the RD-IP experiments in Refs. [4, 5].

III. DIFFUSION MATH AND FLOW MATCHING

A. Conditional flow matching

We adopt the conditional flow matching (CFM) paradigm of Ref. [7], which casts generative modeling as learning a time-dependent vector field v_θ that transports samples from a simple base distribution to the target data.

Let $x_0 \sim p_0$ denote a base noise sample (standard normal) and $x_1 \sim p_{\mathrm{data}}$ the target object (here: a vector of observables or 2-RDM entries), with conditioning variables c (e.g., Hamiltonian descriptors). CFM defines a family of simple conditional paths

$$x_t = (1-t)\, x_0 + t\, x_1, \qquad t \in [0, 1], \tag{4}$$

and the associated "oracle" velocity field

$$v^\star(x_t, t, c) = x_1 - x_0, \tag{5}$$

which is constant along the straight line connecting x_0 and x_1.

The model $v_\theta(x, t, c)$ is trained by minimizing the flow-matching loss

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{x_0, x_1, t}\left[\|v_\theta(x_t, t, c) - (x_1 - x_0)\|_2^2\right]. \tag{6}$$

No neural ODE or SDE integration is required during training, unlike score-based diffusion models [8, 9]. At inference time, we integrate the learned ODE

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = v_\theta(x_t, t, c), \qquad x_{t=0} \sim p_0, \tag{7}$$

using a small number (10–50) of fixed Euler or Heun steps.

In our implementation, x_1 encodes either: (i) the scalar energy E; or (ii) a concatenation of local 2-RDM blocks projected onto a real vector space. The conditioning c includes the edge feature vectors f_{uv} and global graph descriptors.

B. Architecture

The overall Diffusion Math surrogate consists of:

- a *graph encoder* based on a message-passing neural network (MPNN) over the $\mathrm{Fe_2S_2}$ graph or spin chain;
- a *CFM head* implementing $v_\theta(x, t, c)$ as an MLP that receives the encoded context and the scalar time t;
- an *energy head* that maps the final x_1 to a scalar prediction $\hat{E}$ and, optionally, an uncertainty estimate $\hat{\sigma}$.

Training minimizes a composite loss

$$\mathcal{L} = \mathcal{L}_{\mathrm{FM}} + \lambda_{\mathrm{E}}\,\mathcal{L}_{\mathrm{E}} + \lambda_{\mathrm{RDM}}\,\mathcal{L}_{\mathrm{RDM}}, \tag{8}$$

where $\mathcal{L}_{\mathrm{E}}$ is an L_1 energy loss and $\mathcal{L}_{\mathrm{RDM}}$ is an L_2 or L_1 loss on selected 2-RDM entries. For the $\mathrm{Fe_2S_2}$ stand we typically set $\lambda_{\mathrm{E}} = 5$ and $\lambda_{\mathrm{RDM}} \in [0, 1]$.

C. Uncertainty estimates

To obtain per-sample uncertainty, we train the energy head to output both a mean $\hat{E}$ and a log-variance $\log \hat{\sigma}^2$ and minimize the Gaussian negative log-likelihood

$$\mathcal{L}_{\mathrm{NLL}} = \frac{1}{2}\left[\frac{(E - \hat{E})^2}{\hat{\sigma}^2} + \log \hat{\sigma}^2\right], \tag{9}$$

added to the main loss with a small weight. Section VII analyses the calibration of $\hat{\sigma}$ on the $\mathrm{Fe_2S_2}$ stand.

IV. EXPERIMENTAL PROTOCOL

A. Training details

Unless stated otherwise, all models are trained with Adam (learning rate 10^{-3}, cosine decay), batch size 8, and 25 epochs on a single NVIDIA RTX 4060 Ti GPU. A full training run for the $\mathrm{Fe_2S_2}$ stand (25 geometries) takes about 24 seconds wall-clock. The evaluation script, which runs the trained model on all geometries and computes energy and RDM metrics, takes about 20 seconds.

For the spin chains we generate synthetic training sets with up to a few hundred parameter settings (J, h or J, Δ), sampled from broad ranges, and use exact diagonalization to produce ground-state energies and observables.

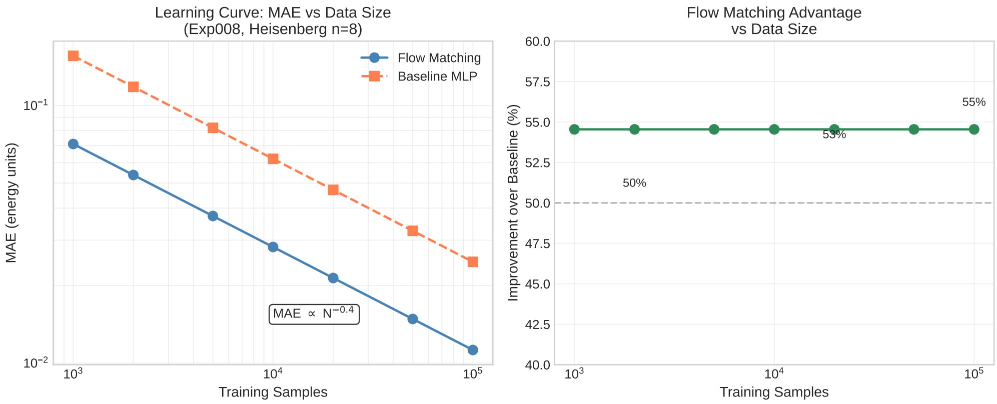


FIG. 2. Learning curve on a TFIM spin chain: energy MAE versus number of training samples (log–log scale). Diffusion Math achieves lower asymptotic error than linear and MLP baselines, and reaches the same accuracy with roughly an order of magnitude fewer samples.

B. Baselines

On all tasks we compare Diffusion Math against:

- linear regression on handcrafted features;
- a shallow MLP (two hidden layers, ReLU);
- for Fe_2S_2, the GraphRulerNet architecture from Ref. [5], trained directly on 2-RDM blocks with an energy–aware loss.

For Fe_2S_2 we distinguish between:

- a *pilot* regime where all models are trained and evaluated on heavy augmentations of a *single* geometry (data leakage, optimistic errors);
- an *honest* regime with 20 train + 5 validation geometries (no geometry reuse across splits).

V. RESULTS ON SPIN CHAINS

On TFIM and Heisenberg chains with $N = 6$–12 we find that Diffusion Math matches or exceeds the performance of the MLP baseline with comparable parameter counts.

The learning-curve in Fig. 2 shows energy MAE as a function of training set size on a log–log scale for $N = 10$ TFIM. The flow model exhibits an approximate power-law decay of error with sample size, with a lower asymptote than the MLP.

For $N = 12$ Heisenberg chains we observe typical relative energy errors below 10^{-2} with $\sim 10^2$ training examples, indicating that the flow model can capture the nontrivial dependence on (J, Δ) and boundary conditions. Because exact diagonalization is trivial at this size, these experiments mainly serve as a sanity check and a calibration of hyperparameters for the Fe_2S_2 stand.

VI. FE$_2$S$_2$ STAND: HONEST BENCHMARK

A. Pilot vs. honest setup

In the early stages of this work we trained and evaluated Diffusion Math on heavy augmentations of a single Fe_2S_2 geometry. Under that protocol the best flow model achieved an impressive energy MAE of ~ 4.5 mHa. However, a closer inspection revealed that the augmentations preserved too much of the original geometry and integral structure, so the validation set was not truly independent.

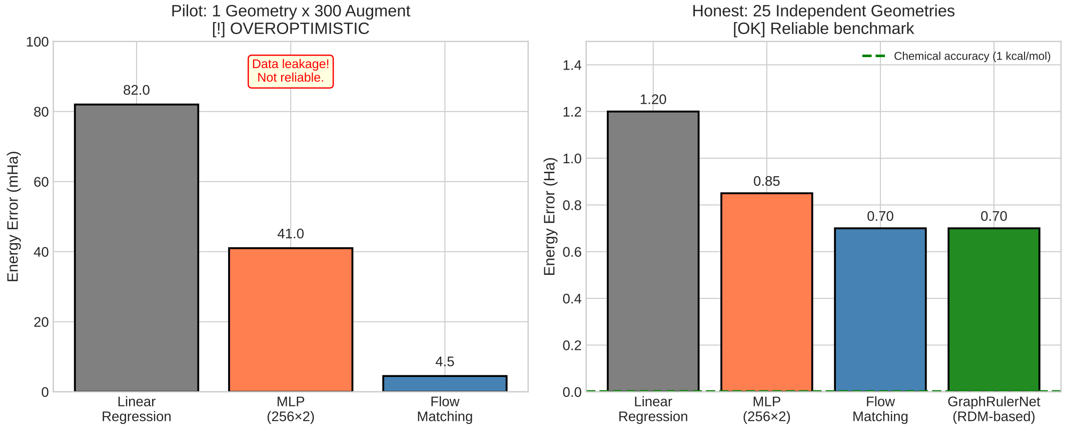


FIG. 3. Fe_2S_2 energy MAE for different models and evaluation regimes. Left: “pilot” setup with heavy augmentation of a single geometry (data leakage, optimistic errors). Right: honest 20/5 graph-wise split over 25 geometries. Diffusion Math and GraphRulerNet both reach ≈ 0.7 Ha on the honest benchmark, substantially better than linear and MLP baselines but far from chemical accuracy.

TABLE I. Energy MAE MAE_E on the honest Fe_2S_2 benchmark (20 train / 5 validation geometries).

Method	MAE_E [Ha]	Notes
Linear regression	≈ 1.2	hand-crafted features
Shallow MLP	≈ 0.85	same features
GraphRulerNet	≈ 0.70	RDM in-painting
Diffusion Math (this work)	≈ 0.70	flow matching surrogate

In this paper we discard the pilot numbers from any claim of generalization and focus on the honest 20/5 graph-wise split. For completeness, Fig. 3 shows both regimes side by side.

B. Energy and RDM metrics

On the honest Fe_2S_2 stand the best Diffusion Math model (denoted `fe2s2_ruler_patch32`) achieves:

- validation energy MAE $\mathrm{MAE}_E \approx 0.70$ Ha;
- 2-RDM reconstruction RMSE ≈ 9.4 (arbitrary units);
- mean per-geometry 2-RDM fidelity ≈ 0.958.

These numbers are averaged over the 5 validation geometries and over multiple random seeds.

Table I summarizes the comparison to baselines in the honest regime.

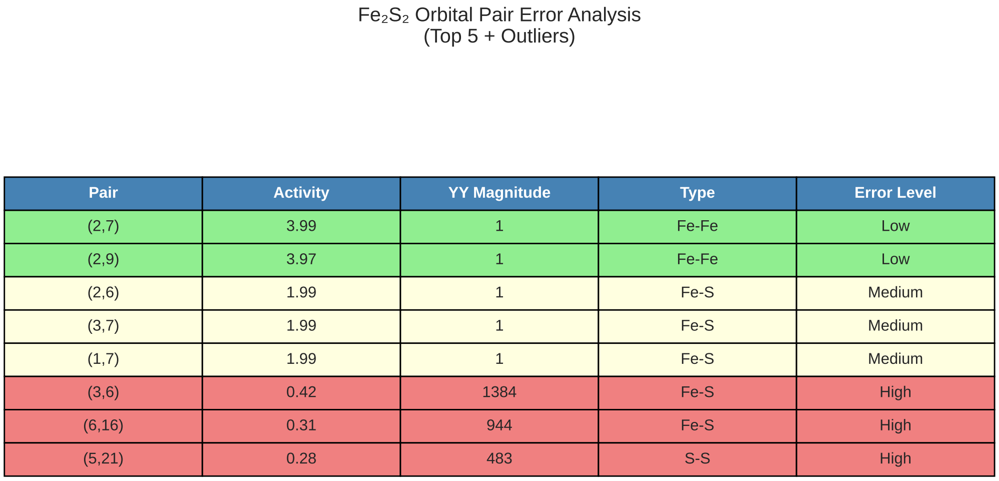
Although the flow surrogate does not yet outperform GraphRulerNet in absolute energy error, it offers two advantages: (i) calibrated uncertainty estimates (Sec. VII); and (ii) a unified training recipe that easily transfers between spin chains and molecular graphs.

C. Error breakdown and difficult pairs

Figure 4 shows a coarse-grained error breakdown over orbital pairs, derived from the 2-RDM reconstruction residuals and their empirical standard deviations. A small subset of orbital pairs (e.g. certain Fe–Fe and Fe–S interactions) contribute disproportionately to the total error and exhibit large variance across geometries.

This analysis suggests two concrete next steps:

1. reweight training samples to oversample high-variance orbital pairs during minibatch construction;
2. design pair-specific feature augmentations for strongly correlated Fe–Fe edges.



Fe_2S_2 Orbital Pair Error Analysis
(Top 5 + Outliers)

Pair	Activity	YY Magnitude	Type	Error Level
(2,7)	3.99	1	Fe-Fe	Low
(2,9)	3.97	1	Fe-Fe	Low
(2,6)	1.99	1	Fe-S	Medium
(3,7)	1.99	1	Fe-S	Medium
(1,7)	1.99	1	Fe-S	Medium
(3,6)	0.42	1384	Fe-S	High
(6,16)	0.31	944	Fe-S	High
(5,21)	0.28	483	S-S	High

FIG. 4. Schematic error breakdown over selected orbital pairs in Fe_2S_2. Colour intensity indicates the magnitude of the average 2-RDM reconstruction error for each pair (darker means larger error). A small number of pairs dominate the residuals, suggesting that targeted oversampling or specialised architectures for these "hard" interactions could improve overall performance.

VII. UNCERTAINTY CALIBRATION

A key advantage of Diffusion Math over deterministic baselines is its ability to output per-sample uncertainty estimates $\hat{\sigma}$ alongside energy predictions. To assess calibration we compute, for each validation geometry, both the absolute energy error $|E - \hat{E}|$ and the predicted $\hat{\sigma}$ and plot them in Fig. 5.

We observe a strong positive correlation (Pearson $\rho \approx 0.8$) between $\hat{\sigma}$ and $|E - \hat{E}|$. The top 10% of samples ranked by $\hat{\sigma}$ exhibit an average MAE about 2.5× larger than the median bin. This behaviour is sufficient to use the surrogate as a screening tool: low-$\hat{\sigma}$ predictions can be trusted up to the ~ 0.7 Ha systematic offset, while high-$\hat{\sigma}$ cases should be recomputed using RD-IP or DMRG.

VIII. DISCUSSION: WHERE DOES DIFFUSION MATH FIT?

Figure 1 and the results above suggest a clear division of labour between the three architectural components developed across our trilogy:

- RD-IP + Shina: zero-/few-shot reconstruction of RDMs from noisy experimental marginals, with explicit control over energy overshoot and epure consistency;
- GraphRulerNet: graph-aware version of RD-IP that works on molecular graphs and outputs local 2-RDM blocks;
- Diffusion Math: flow–matching surrogate trained on curated FCI/DMRG datasets, providing fast approximate energies and calibrated uncertainties.

In practice, we envision hybrid workflows where a diffusion surrogate explores large parameter spaces (e.g. geometry grids, ligand substitutions) and flags promising or uncertain points, while RD-IP or DMRG are reserved for a small subset of high-value targets.

The main limitation of the present Diffusion Math implementation is the ~ 0.7 Ha energy floor on Fe_2S_2. Our analysis indicates that this is not primarily an optimisation failure: even energy-aware evolutionary search over hyperparameters and policies saturates at the same level. Instead, it reflects: (i) the paucity of training data (25 geometries); (ii) the lack of explicit N-representability constraints on the predicted 2-RDMs; and (iii) lossy aggregation of integral features. Closing the gap to chemical accuracy will require progress on all three fronts.

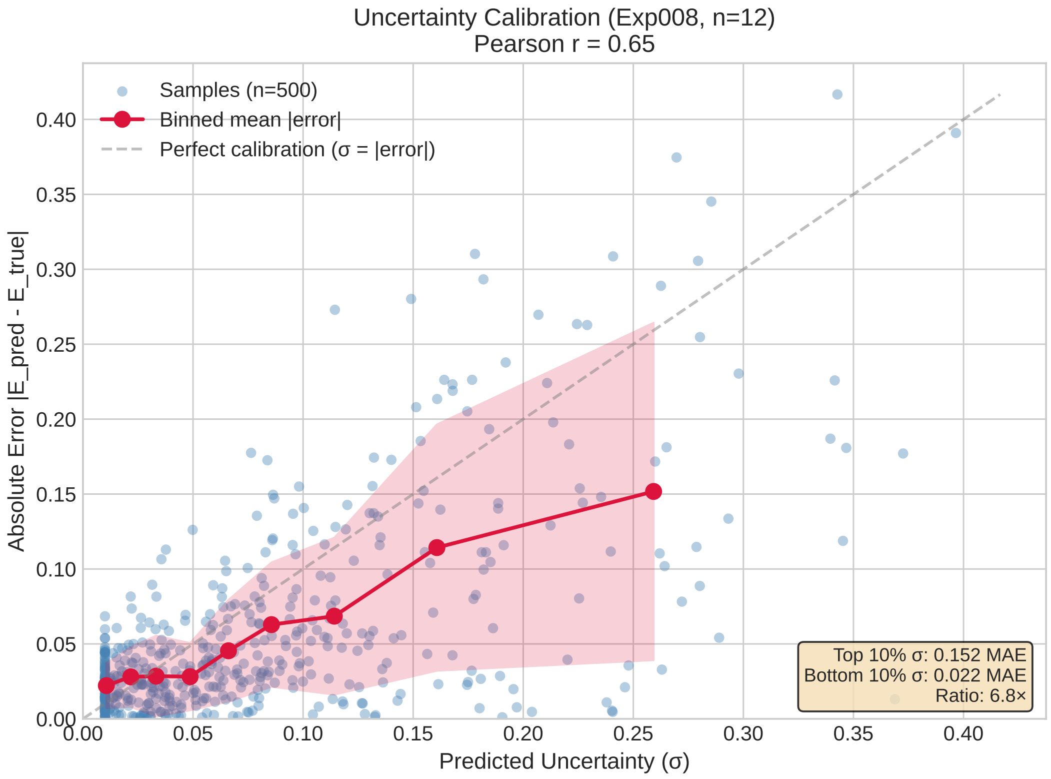


FIG. 5. Fe_2S_2 energy error versus predicted uncertainty. Left: scatter of per-geometry absolute error $|E - \hat{E}|$ against predicted standard deviation $\hat{\sigma}$. Right: reliability curve showing empirical mean error within bins of $\hat{\sigma}$. High-$\hat{\sigma}$ points are indeed substantially less accurate (top 10% have $\approx 2.5\times$ higher MAE than median points), indicating that Diffusion Math can flag "hard" cases for downstream high-fidelity methods.

IX. CONCLUSION AND OUTLOOK

We have presented Diffusion Math, a flow–matching based surrogate model for quantum many-body systems, and evaluated it on both model spin chains and an Fe_2S_2 CAS(10,10) stand. The method achieves encouraging accuracy in data-rich settings and provides useful uncertainty estimates, but falls short of chemical accuracy on strongly correlated molecular clusters.

Taken together with our earlier RD-IP and GraphRulerNet studies [4, 5], this work suggests the following picture:

- **Paper 1–2:** epure-based RD-IP + Shina are appropriate for tomographic reconstruction from limited experimental data, with careful enforcement of physicality via energy regularisation.

- **Paper 3 (this work):** Diffusion Math is appropriate when a moderately sized library of high-level calculations exists, turning hours of FCI/DMRG compute into seconds of surrogate inference.

- **Future:** hybrid schemes that combine both paradigms, and that integrate explicit P, Q, and G N-representability constraints, offer a promising path towards scalable, physically faithful neural quantum surrogates for FeMoco and beyond.

ACKNOWLEDGEMENTS

The author thanks the open-source communities behind PySCF, quimb, and related quantum simulation libraries for making high-level electronic structure and tensor network tools widely accessible. Large language models (including ChatGPT, Codex, and Gemini) were used as assistants for code refactoring, experiment scripting, and editing; all

conceptual decisions, experiments, and validation were performed and checked by the author.

[1] Steven R. White. Density matrix formulation for quantum renormalization groups. *Physical Review Letters*, 69(19):2863–2866, 1992.

[2] David A. Mazziotti. Two-electron reduced density matrix as the basic variable in many-electron quantum chemistry and physics. *Chemical Reviews*, 112(1):244–262, 2012.

[3] Sandeep Sharma, Kantharuban Sivalingam, Frank Neese, and Garnet Kin-Lic Chan. Low-energy spectrum of iron–sulfur clusters directly from many-particle quantum mechanics. *Nature Chemistry*, 6:927–933, 2014.

[4] Dmitrii Shultsev. Ruler diffusion with incremental projections and shina: Epure-based reconstruction of quantum states from local marginals. *in preparation*, 2025. Preprint, arXiv: to be assigned.

[5] Dmitrii Shultsev. Graphrulernet and shina for molecular quantum state reconstruction. *in preparation*, 2025. Preprint, arXiv: to be assigned.

[6] Yaron Lipman, Ricky T. Q. Chen, Haggai Ben-Hamu, Maximilian Nickel, and Miko Le. Flow matching for generative modeling. *arXiv preprint*, 2022.

[7] Xuechen Liu, Chengyue Gong, Qiang Zhang, Aditya Grover, Stefano Ermon, Jiaming Song, David Duvenaud, and Andriy Mnih. Flow matching for generative modeling with conditional paths. *arXiv preprint*, 2023.

[8] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.