Supplementary Information

POLARIS: Perovskite Optimization using LLM-Assisted Refinement and Intelligent Screening

Jordan Marshall[1‡], Sheryl L. Sanchez[1‡], Rushik Desai[2], Elham Foadian[1,] Utkarsh Pratiush[1], Arun Mannodi-Kanakkithodi[2], Sergei V. Kalinin[1], Mahshid Ahmadi[1*]

[1] *Institute for Advanced Materials and Manufacturing, Department of Materials Science and Engineering, The University of Tennessee Knoxville, Knoxville, Tennessee 37996, United States*

[2] *School of Materials Engineering, Purdue University, West Lafayette, IN 47907, USA*

*Corresponding authors Email: mahmadi3@utk.edu

‡ These authors contributed equally

Code Availability

The code can be found on the Github repository: https://github.com/SLKS99/POLARIS-workflow-stage-1/tree/main

Step	Prompt Purpose	Example Prompt	Expected Output Format
1	Extract experimental + results sections	Extract the experimental and results sections from this document.	Full text blocks (usually paragraphs)
2	Identify perovskite compounds	List all perovskite compounds mentioned in the Experimental or results sections.	Bullet list or table of perovskite formulas (e.g., $CsPbBr_3$)
3	Extract materials + IUPAC names	For each perovskite, list all materials used in its synthesis and their IUPAC names.	Structured table: Perovskite
4	Add synthesis conditions	List the synthesis temperature, method, and solvent used for each perovskite.	Table with added columns: Temperature, Synthesis Method, Solvent
5	Re-run or adapt prompt	Double check the synthesis info against mentions in the results section for consistency.	Confirmed/revised synthesis conditions

Table S1: Five-step LLM prompting workflow for extracting and validating perovskite synthesis details

Feature	**Polak et al. (Digital Discovery, 2024)** Flexible, Model-Agnostic	**Polak & Morgan (Nature Comm., 2024)** ChatExtract	**This Work:** NotebookLM + Prompt Library with RAG
Preprocessing Format	HTML/XML → Sentence Tokenization	HTML/XML → Sentence Tokenization	Full-text ingestion (PDF/HTML) via browser-based RAG
Segmentation Method	Rule-based sentence splitting	Rule-based sentence splitting	None; full-document context retained
Data Filtering	Regex filters + numeric sentence pruning	Zero-shot sentence classification via LLM	Prompt-driven targeting of experimental sections only
Extraction Method	Single-shot prompts on filtered sentences	Multi-turn LLM with follow-up validation	Structured prompt library + retrieval-augmented memory
Automation Level	Semi-automated with human review	Semi-automated with conversational control flow	Modular and scalable; auto-batch with minimal manual tuning
Triplet Structure (Material–Value–Unit)	Optional; context-dependent	Explicitly enforced with structured formatting	Not enforced; flexible phrase/entity extraction based on use case
Error Mitigation Strategy	Human validation of outputs	Redundant prompts + hallucination rejection	Conservative prompting + source-grounded extractions
LLM Models Used	GPT-3/3.5/4, BART, DeBERTaV3	ChatGPT-3.5/4 (OpenAI)	NotebookLM, GPT-4o, DeepSeek-V3, Publication Analyzer
Token Cost Consideration	Low due to sentence pre-filtering	Moderate to high; mitigated with prompt design	**Not applicable** — All tools used via free or flat-rate web UIs; no API tokens used
Reproducibility	Medium (human-in-the-loop variability)	High (conversational state retention)	Very high (static prompt library, document memory, batch consistency)
Best Use Case	Mid-sized numerical datasets with known formats	Structured triplet databases from materials text	Hypothesis-driven curation from heterogeneous experimental literature

Table S2: Comparative summary of preprocessing formats, segmentation and filtering methods across various LLM-driven materials text-mining frameworks.

Spacer Cation	Structure Type	Key Features	Effects on 2D Halide Perovskite
n-Butylammonium (BA)	Linear alkyl (C4)	Rigid; one F on ring (para position) adding dipole	Higher dipole moment than PEA; induces an internal field in perovskite layers, aiding charge separation. Still hydrophobic (F is slightly hydrophobic substituent). Reported to improve stability and red-shift bandgap slightly by influencing layer interactions. Represents *dipolar aromatic* category.
iso-Butylammonium	Branched alkyl	Two halogens on ring (F and Cl) increasing dipole and polarizability	Strong dipole and heavy-atom substituents can enhance spin–orbit interactions and lattice energy. Used to **stabilize [PbI_6] octahedra** and promote ordered crystal growth.Yields very stable 2D/3D heterostructure films with vertical orientation. Less explored in single-phase 2D, but promising for stability.
2-Phenylethylammonium (PEA)	Aromatic (phenyl–$CH_2CH_2NH_3$)	Rigid; sulfur in ring (heteroatom)	Thiophene is slightly more electron-rich and elongated than benzene. ThEA spacers have been shown to form RP perovskites with good stability; the S heteroatom can engage in weak coordination or enhanced van der Waals with iodine. This spacer expands the library beyond benzene – e.g. $(ThEA)_2PbI_2$has similar bandgap to $(PEA)_2PbI4$but potentially different exciton dynamics. Still largely unexplored in devices.
4-Fluorophenethylammonium (4F-PEA)	Aromatic with electronegative substituent	Rigid alicyclic (chair conformation)	Lacks π-bonding; purely sigma bonds. Very bulky cross-section. Tends to enforce wider layer spacing. Could improve thermal stability due to rigid packing, but the electronic coupling is reduced (large gap). Not commonly used in high-efficiency cells (more in light-emitting applications).
Mixed Halogenated Phenethyl (e.g. 3F-4Cl-PEA)	Aromatic with multi-substituents	Two fused rings (naphthalene); large π system	Very bulky planar cation, highly hydrophobic. Forms wider interlayer spacing; increased aromatic stacking interactions. Has been used to create wider quantum wells with red-shifted absorption. However, the size can lead to phase segregation or slow crystallization. Still a part of the *explored* space, but devices often show lower current due to wider bandgap (good for tandem or color-tunable cells).
Thienylethylammonium (ThEA)	Heterocyclic aromatic (thiophene ring)	Symmetric, delocalized charge; high polarity	Thienylethylammonium (ThEA) as a spacer cation in 2D Ruddlesden–Popper perovskites introduces a conjugated thiophene ring that increases organic-layer polarizability and reduces dielectric confinement,

			leading to lower exciton binding energies compared with aliphatic ammonium spacers. The delocalized π system and high polarity of ThEA promote stronger electronic coupling at the organic–inorganic interface and can improve interlayer charge transport. The aromatic heterocycle also supports π–π interactions between spacer layers, contributing to lattice stabilization and improved environmental robustness, while its steric profile influences octahedral tilting and slab spacing, subtly tuning band structure and optical emission.
Cyclohexylammonium (CHA)	Cyclic aliphatic	Rigid, no π electrons; can have substituents like –F	These have been recently explored (short-chain *N*-butylpiperidinium with F) for photodetectors. The ring provides a fixed cationic scaffold, and substituents like –F break symmetry to add a dipole. They form stable 2D perovskites with low bandgaps (when mixed with 3D components). So far used in niche cases – represent a new direction for spacers combining rigidity and dipolar design.
1-Naphthylethylammonium (NEA)	Polycyclic aromatic	e.g. *meta*-substituted aromatics (to tune packing angle), amino-acid-based ammoniums (NH_3+CH(R)–COO zwitterions), or bulky cations with internal hydrogen bonding	These areas remain open. **Meta-substituted aromatics** (as opposed to para) could lead to different layer tilts due to angle of substituent. **Zwitterionic spacers** (with internal COO^- group balancing NH_3^+ might pack differently and offer strong binding to Pb^{2+} (though risk breaking perovskite network). **Bulky cations with intramolecular H-bonds** (to lock conformation) could yield very well-defined spacing. Each of these is largely theoretical so far, highlighting the expansive design space beyond the common linear or phenyl spacers.
Guanidinium (Gua) – *Note:* divalent cation	Planar polyamine $(C(NH_2)^{3+}$	Symmetric, delocalized charge; high polarity	*Not a RP spacer in the monovalent sense*, but worth noting: guanidinium can sometimes insert between layers or in the A-site of 3D perovskites. It has a large size and planar shape, but due to double positive charge would form Dion–Jacobson phases if used as primary spacer. Guanidinium-like spacers remain largely **unexplored** in true RP structures (requiring charge balance adjustments), but could offer strong hydrogen bonding (multiple –NH_2) and very large dipole.
Piperidinium derivatives	Cyclic amine	Rigid, no π	These have been recently explored

(e.g. fluorinated piperidinium)	(six-membered ring, NH in ring)	electrons; can have substituents like –F	(short-chain *N*-butylpiperidinium with F) for photodetectors. The ring provides a fixed cationic scaffold, and substituents like –F break symmetry to add a dipole. They form stable 2D perovskites with low bandgaps (when mixed with 3D components). So far used in niche cases – represent a new direction for spacers combining rigidity and dipolar design.
Unexplored novel cations	–	e.g. *meta*-substituted aromatics (to tune packing angle), amino-acid-based ammoniums (NH_3^+ CH(R)–COO^- zwitterions), or bulky cations with internal hydrogen bonding	These areas remain open. **Meta-substituted aromatics** (as opposed to para) could lead to different layer tilts due to angle of substituent. **Zwitterionic spacers** (with internal COO^- group balancing NH^{3+} might pack differently and offer strong binding to Pb^{2+} (though risk breaking perovskite network). **Bulky cations with intramolecular H-bonds** (to lock conformation) could yield very well-defined spacing. Each of these is largely theoretical so far, highlighting the expansive design space beyond the common linear or phenyl spacers.

Table S3: Organic Spacer Suggestions from ChatGPT

Cluster	Representative Chemistry	Key Distinguishing Features (from Figures S3–S5)
0	Long linear alkyl ammoniums	Lowest TPSA, lowest polarity, largest logP; weak electronic coupling; good moisture barrier.
1	Bulky π-conjugated aromatics (e.g., naphthalene, phenyl variants)	High HOMO (shallower), moderate dipole; strong π–π interactions; wider organic layers.
2	Small cyclic aliphatic spacers	Compact, rigid, low conjugation; dense packing; suppress phase transitions.
3	Small allylic/amino vinyl groups	Very small cations; introduce slight π character; good for tuning orientation.
4	Diammonium / dithiol bifunctional species	Highest TPSA, strongest donor/acceptor counts; strong binding to Pb sites; best defect passivation.
5	Halogenated alkyl ammoniums	Intermediate polarity; halogen substitution modulates HOMO–LUMO and dielectric environment.

Table S4: Clusters and the representative features

Note 1: Benchmarking Metrics Calculation

Ground truth phrase recall (%) was calculated as the number of correctly extracted multi-word gold-standard phrases divided by the total number of gold-standard phrases, multiplied by 100. This metric evaluates whether the model is able to return complete, semantically meaningful chunks exactly as defined in the annotated reference set. In contrast, ground truth token recall (%) measures the number of individual tokens (words or subwords) within the gold-standard phrases that were captured in the model output, divided by the total number of ground-truth tokens, also multiplied by 100. This allows for partial credit in cases where the model did not reconstruct an entire phrase but still recovered some of its lexical content. Ground truth token precision (%) was defined as the number of correct tokens output by the model divided by the total number of tokens it generated. The harmonic mean of token precision and recall yields the GT token F1 score (%), which balances over-generation and under-generation tendencies. Together, these metrics quantify exact phrase-level recovery, partial lexical overlap, and the trade-off between recall and precision, providing a comprehensive evaluation of how well models retrieve and structure key scientific content.

Ground Truth Phrase Recall (%)

$$\left(\frac{\#\ of\ correctly\ extracted\ GT\ phrases}{\#\ of\ total\ GT\ phrases}\right) \times 100$$

Ground Truth Token Recall (%)

$$\left(\frac{\#\ of\ correctly\ matched\ GT\ tokens}{\#\ of\ total\ GT\ tokens}\right) \times 100$$

Ground Truth Token Precision (%)

$$\left(\frac{\#\ of\ correctly\ matched\ GT\ tokens}{\#\ of\ tokens\ output\ by\ model}\right) \times 100$$

GT Token F1 Score (%)

$$2 \times \left(\frac{Precision \times Recall}{Precision + Recall}\right)$$

Where precision and recall are the GT Token Precision (%) and GT Token Recall (%), respectively. These metrics are applied per model output against a predefined set of annotated answers (ground truth) for each prompt, and then aggregated.

Verbosity Ratio is calculated as:

$$\frac{\#\ of\ tokens\ in\ model\ output}{\#\ of\ tokens\ in\ the\ ground-truth\ answer(s)}$$

This metric quantifies how *verbose* the model is relative to the expected answer length. A verbosity ratio close to 1 means the model's output is approximately the same length as the reference. Values significantly above 1 indicate that the model is more verbose than necessary, while values below 1 suggest overly terse responses that may omit information. It does not evaluate correctness directly, but it contextualizes recall and precision especially when models overgenerate or undergenerate. This verbosity ratio is reported in Figure S1c and is interpreted in conjunction with ground-truth phrase recall.

Coefficient of variation (CV) was calculated as the standard deviation of the metric value across repeated runs divided by the mean value for that model ($CV = \sigma / \mu$), and is used to quantify run-to-run stability.

Note 2: Model implementation details

The graph neural network encoder operates on molecular graphs with node-, edge-, and graph-level features. Node features have dimensionality in_node_dim, which varies by dataset but typically ranges from 30 to 50 atomic descriptors per node. Edge features have a fixed dimensionality of 11 and encode bond-specific information such as bond type, bond order, and aromaticity. In addition, each molecular graph includes a set of global descriptors of dimension global_dim, typically comprising 46 features including ring statistics, molecular weight, topological polar surface area, and LogP.

The input node features are first projected into a fixed-dimensional latent space using a fully connected linear embedding layer. This layer applies to a linear transformation from in_node_dim to a hidden dimension of 256, producing the initial node embeddings that are passed to subsequent graph convolution layers.

The encoder then applies four successive graph convolution layers based on the Graph Isomorphism Network with Edge features (GINEConv). All GINEConv layers maintain a constant hidden dimensionality of 256 and explicitly incorporate the 11-dimensional edge feature vectors during message passing. Within each GINEConv layer, node representations are updated using a two-layer multilayer perception. The first layer applies a linear transformation followed by a ReLU activation, and the second layer applies a linear transformation without an activation function. Messages are aggregated by summing transformed representations from neighboring nodes, where each message depends on the sum of the source node embedding, the neighbor node embedding, and the corresponding edge feature vector. The aggregated messages are passed through a ReLU activation and regularized using dropout with a probability of 0.20 during training.

After the final GINEConv layer, node-level embeddings are aggregated into a single graph-level representation using global mean pooling. This operation computes the average of all node embeddings in the graph, resulting in a fixed-dimensional molecular embedding of size 256. This pooled representation serves as the latent molecular descriptor used by downstream components of the model.

Task	RBF R^2	Matern R^2	RBF RMSE	Matern RMSE
band_gap	0.944	0.735	0.261	0.569
homo	0.843	0.640	0.270	0.408
lumo	0.927	0.766	0.314	0.563
dipole_moment	0.674	0.382	0.916	1.262
TPSA	0.470	0.852	15.404	8.134
n_rings	0.943	0.945	0.096	0.094
n_hbd	0.985	0.520	0.102	0.579
n_hba	-0.223	0.757	0.948	0.422
n_conj_bonds	0.952	0.900	0.572	0.822
molecular_weight	-0.008	0.742	12.808	6.477

Table S5: Kernel Comparison on model predictability

To confirm the kernel choice of the model RBF and Matern Kernels were performed to evaluate which kernel model would perform better on the prediction tasks shown in table S5. The Median R^2 of the RBG and Matern Kernel are 0.927 and 0.757 showing that the original RBF Kernel has 6 tasks with a R^2 of 0.8 or higher. While Matern has 5 tasks with a R^2 with 0.6 or higher. Although not all tasks performed better with RBF such as TPSA, molecular weight and n_hba. In this case the choice between kernels should be task specific. For electronic property prediction RBF may be preferred. For general-purpose multi-task learning matern may provide a more consistent performance. In this case a model with a hybrid kernel will be used to try to improve the overall performance

Note 3: Ground-Truth Benchmark Paper Selection

For ground-truth benchmarking, ten papers were selected from the full 200-paper corpus through a manual random screening procedure. Papers were chosen by iterating through the collected PDF dataset without targeting any specific journal, author group, or reported outcome, with the sole inclusion requirement being the presence of a clearly identifiable Methods/Experimental section to enable consistent annotation of synthesis-relevant metadata. To ensure heterogeneity in reporting formats and document structure, each of the ten benchmark papers originated from a different journal. This random, cross-journal sampling strategy provides a representative evaluation of extraction performance across diverse publication styles while maintaining a consistent basis for manual ground-truth comparison.

Handling of data fidelity and imbalance

No explicit weighting scheme was applied to differentiate experimentally validated spacer cations from augmented QM9-derived molecules during training. This choice was deliberate. Experimentally reported spacers serve as anchors for the chemically relevant region of latent space and are enforced at the initialization stage by requiring their inclusion in the seed training set. After initialization, all molecules are treated uniformly to avoid introducing subjective or poorly calibrated fidelity weights.

Importantly, numerical imbalance does not imply dominance in prediction. Influence in the GNN–GP model is governed by proximity in the learned latent chemical space rather than by sample count. While QM9-derived molecules are more numerous, they predominantly occupy regions of

chemical space that are distant from literature-derived spacers. In the Gaussian process regression stage, contributions from such distant points are naturally attenuated by the kernel, while predictions near experimentally relevant spacers are dominated by nearby latent representations. The GP further propagates uncertainty in sparsely supported regions, ensuring that augmented data increase epistemic uncertainty rather than overwhelming experimentally anchored predictions. The augmented molecules therefore function primarily as regularizers for representation learning rather than as competing sources of ground truth.

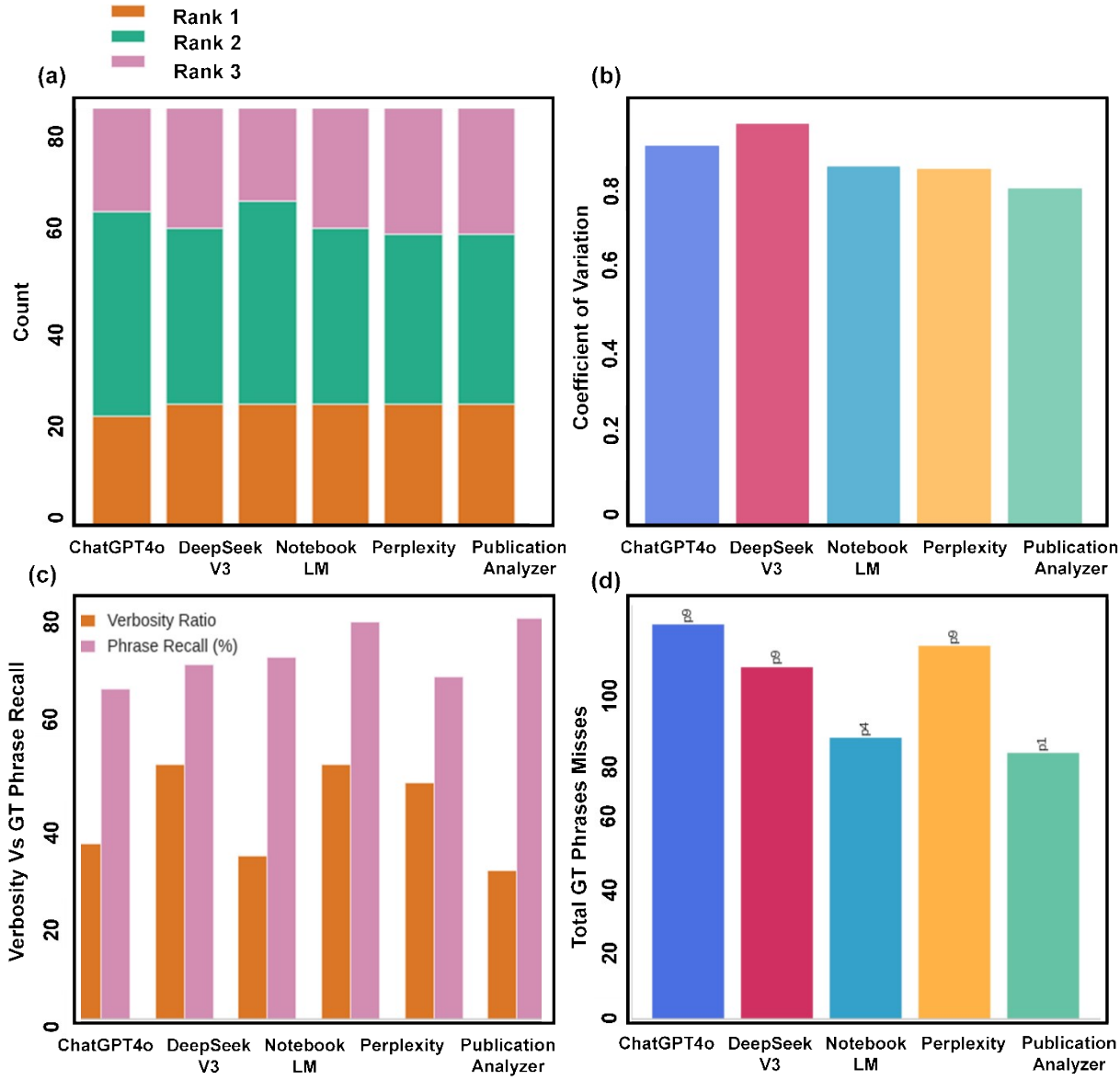

Figure S1: Comparative evaluation of six LLM models across 10 papers: (a) distribution of first-, second-, and third-place rankings by GT token F1 score across three independent trials, (b) coefficient of variation of GT token F1 scores across trials, calculated as

the standard deviation divided by the mean (CV = σ/μ), (c) average verbosity ratio (defined as the ratio of output token count to ground-truth token count; see Note 1) plotted against ground-truth phrase recall, and (d) total number of missed ground-truth phrases aggregated across all benchmark papers, where numerical annotations above each bar indicate the total miss count and letter labels denote the worst-performing individual paper for that model.

DFT Methodology

We utilized the AiiDA[1,2] framework to automate the task of molecular geometry optimization and dipole calculations. SMILES strings were converted to atoms objects and a cheap force field from RDKit[3] was used to confirm the molecular conformer. We utilized a foundational Machine Learning Interatomic Potential (MLIP) called MACE-OFF[4], which is trained on a large dataset of organic molecules, to relax the geometry of the molecules before running additional density functional theory (DFT) computations to obtain the ground state structures. To adequately account for the van-der Waals interactions in the organic species, we used the Projector Augmented Wave[5,6] based Perdew-Bercke-Ernzerhof (PBE)[7] functional within the Generalized Gradient Approximation (GGA) with the D3 correction method of Grimme et. al[8,9]. We used a molecule-in-a-box approach and relaxed the geometry with a kinetic energy cutoff of 500 eV at the G-point using the Vienna Ab-initio Simulation Package (VASP)[10–12]. Using the relaxed geometry, static electronic structure calculations were performed to obtain the dipole moment and HOMO-LUMO gap.

Model information:

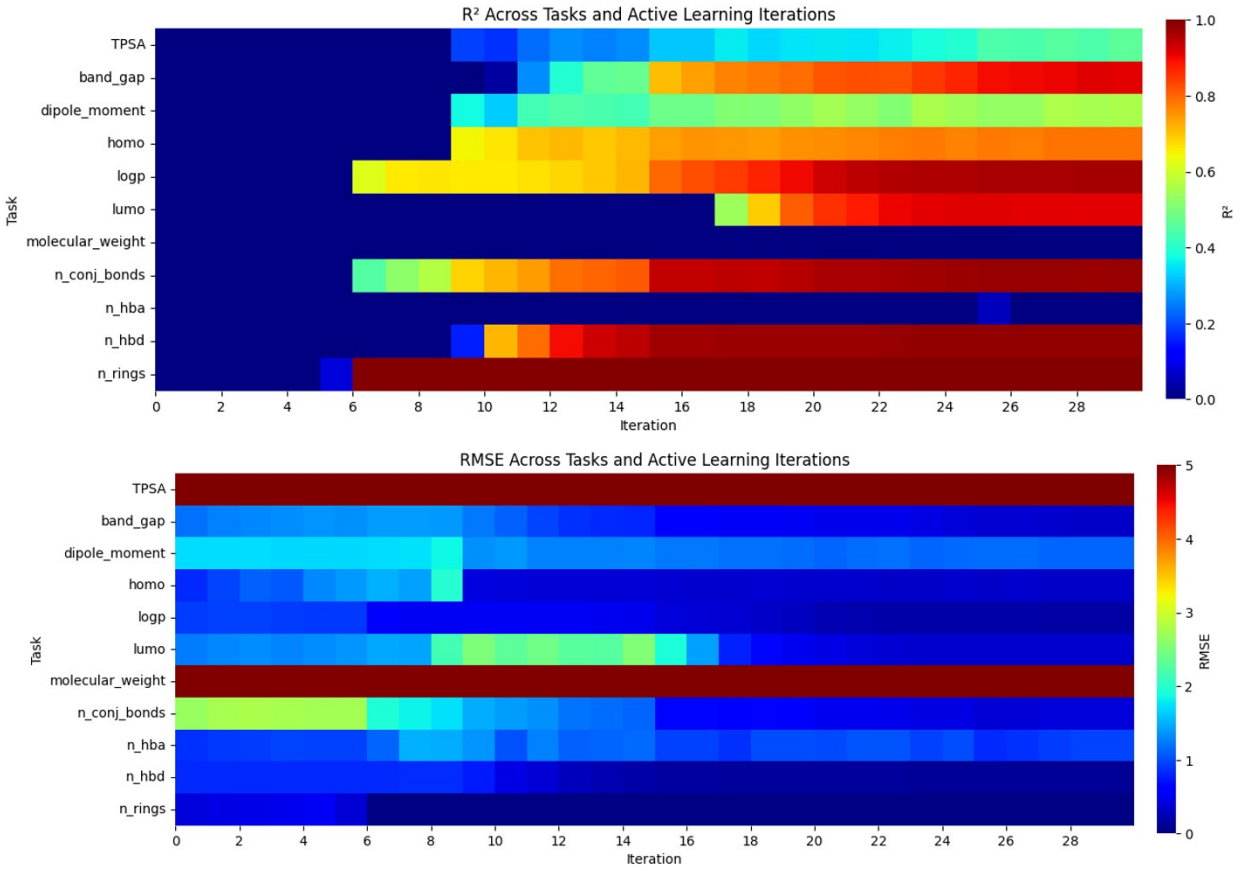

Figure S2: Predictive performance across active learning iterations for multi-task property prediction. Top: coefficient of determination (R^2) per task over 30 active learning rounds. Bottom: root mean squared error (RMSE) over the same iterations. Each row corresponds to a molecular property predicted by a deep kernel learning model with a shared GNN encoder and task-specific Gaussian processes.

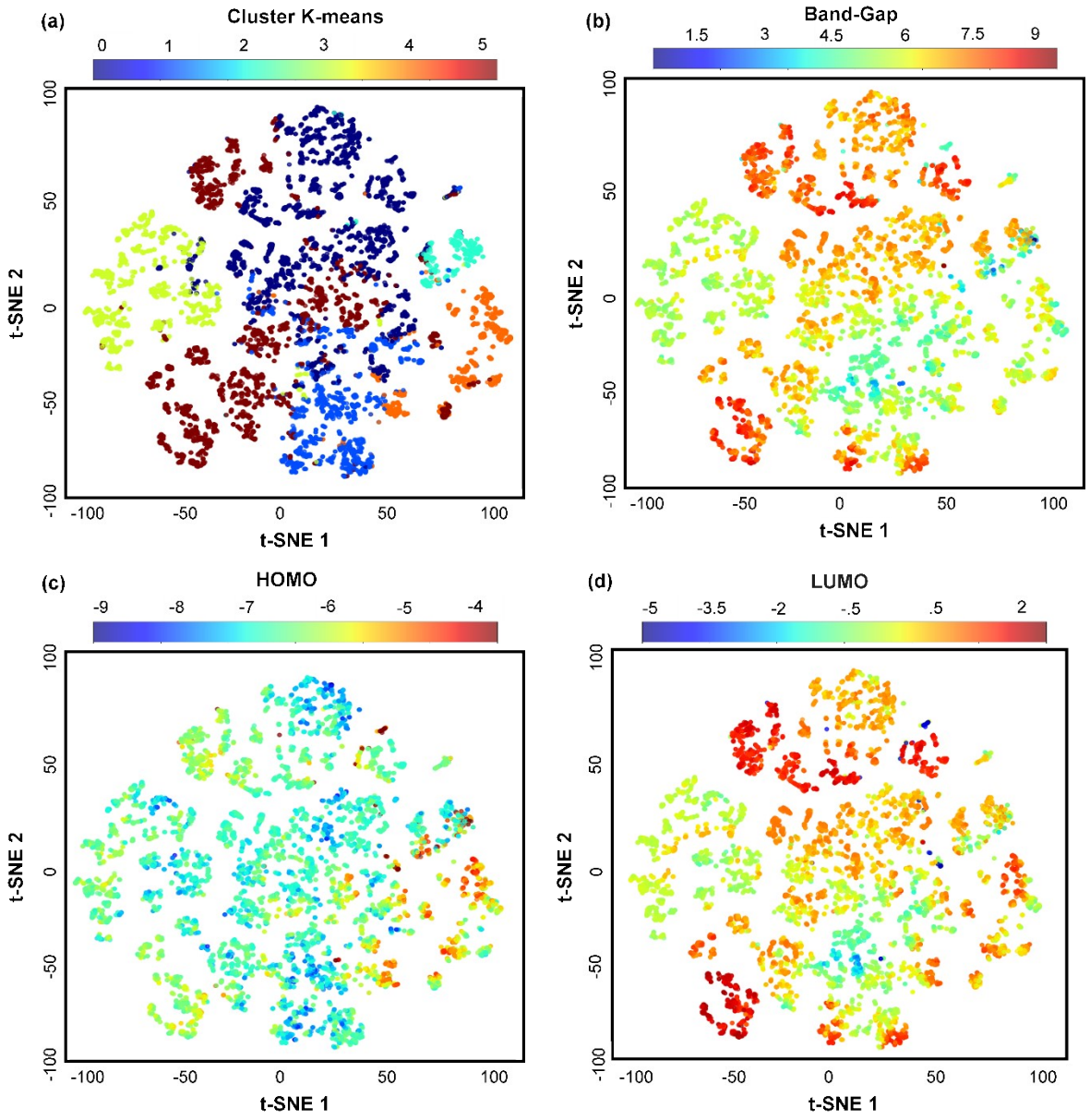

Figure S3: Latent space with t-SNE reduction showing coloring based on (a) k-means clustering, (b) Band gap, (c) HOMO, (d) LUMO.

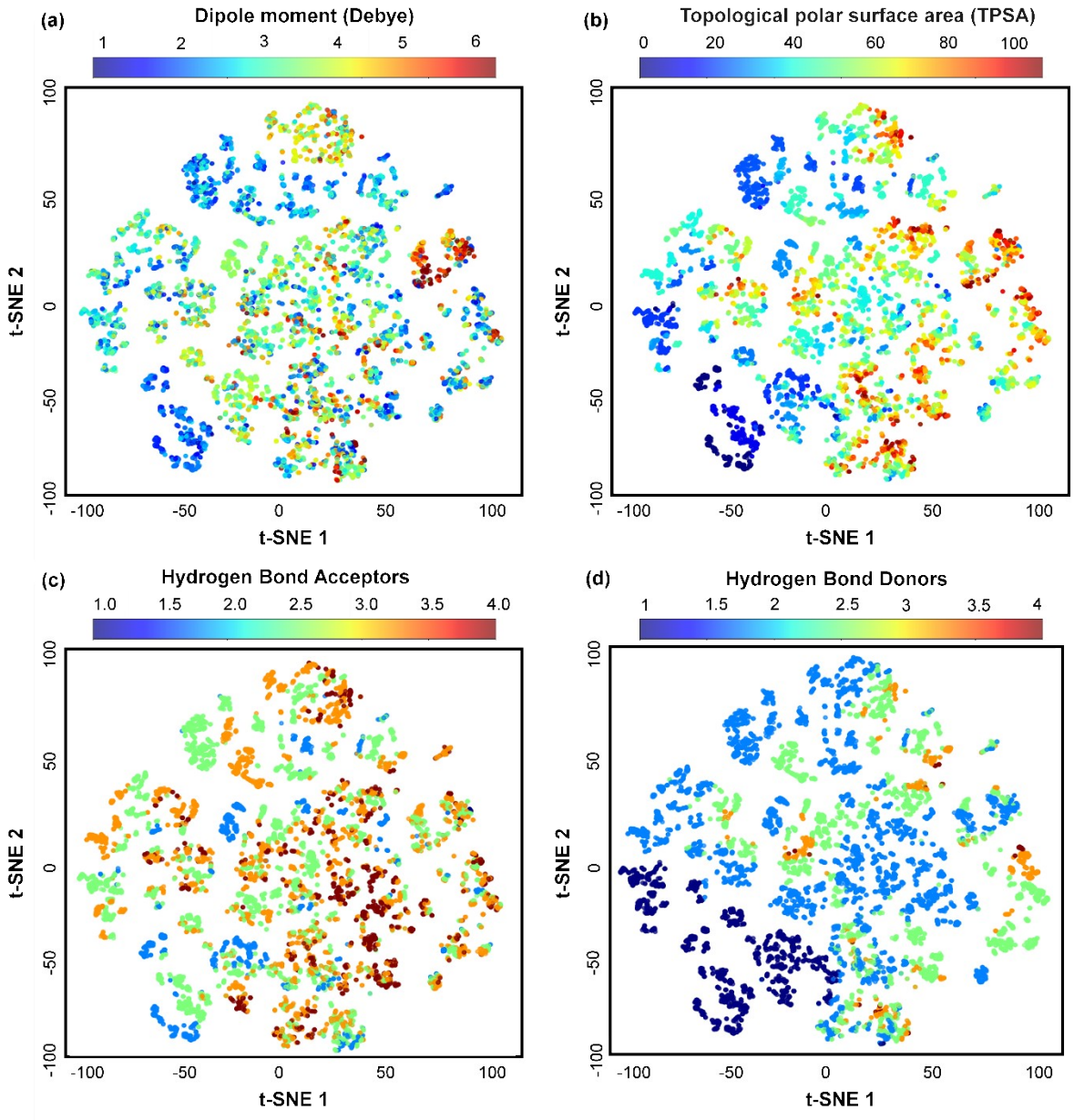

Figure S4: Latent space with t-SNE reduction showing coloring based on (a) Dipole moment, (b) TPSA, (c) Hydrogen bond acceptors, (d) Hydrogen bond donors.

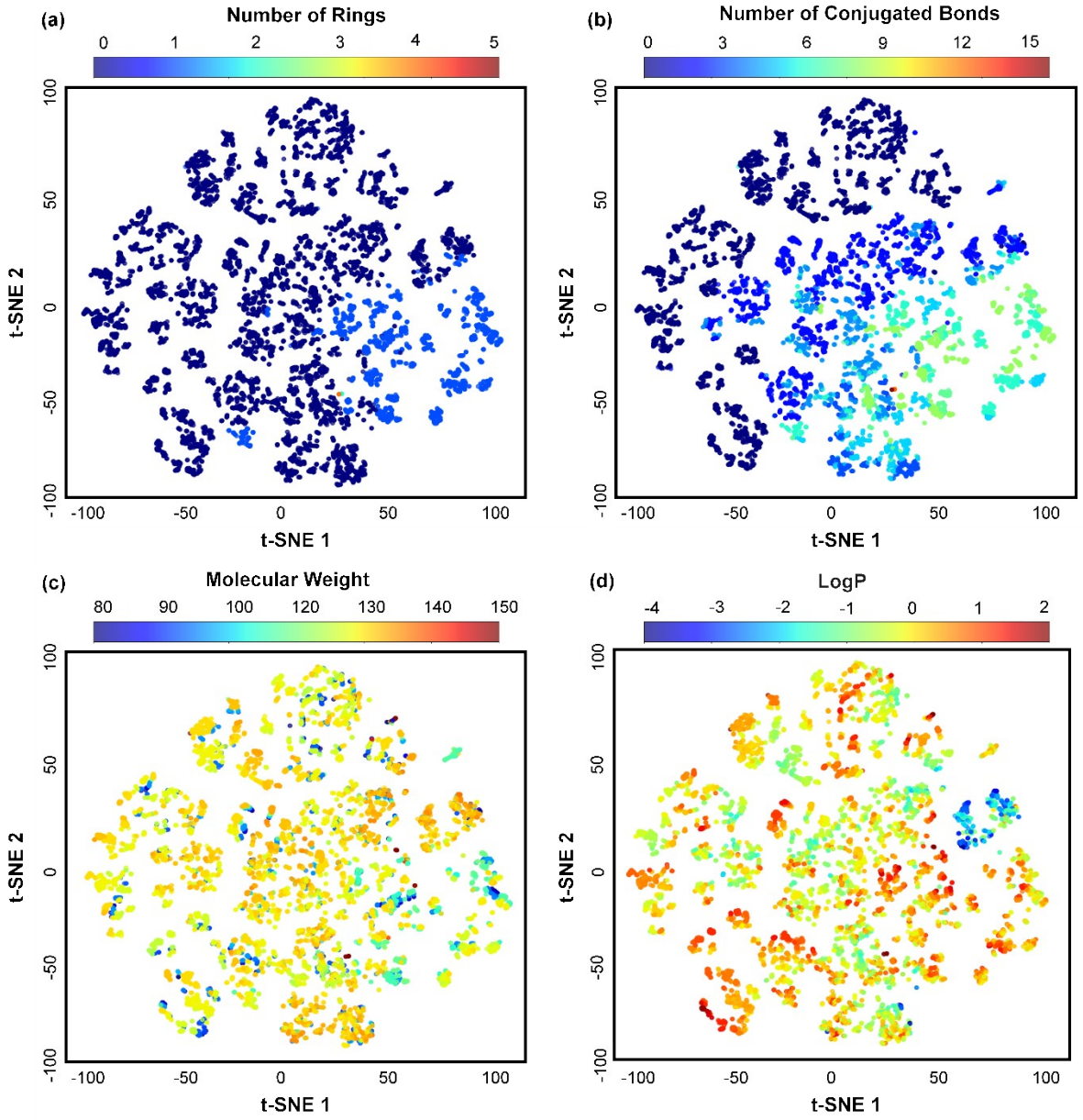

Figure S5: Latent space with t-SNE reduction showing coloring based on (a) number of rings, (b) number of conjugated bonds, (c) molecular weight, and (d) LogP.

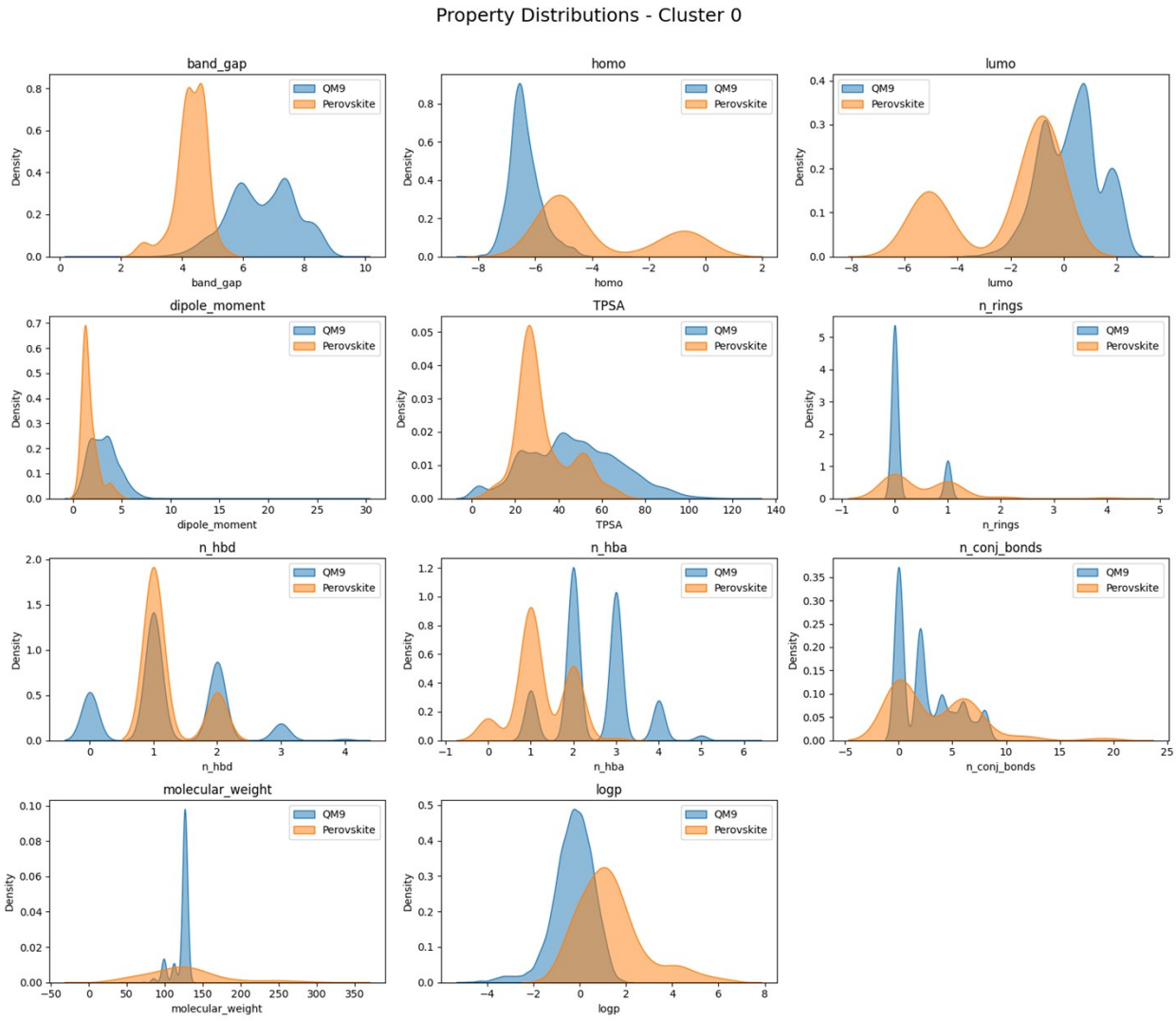

Figure S6: Property distributions for each target for cluster 0.

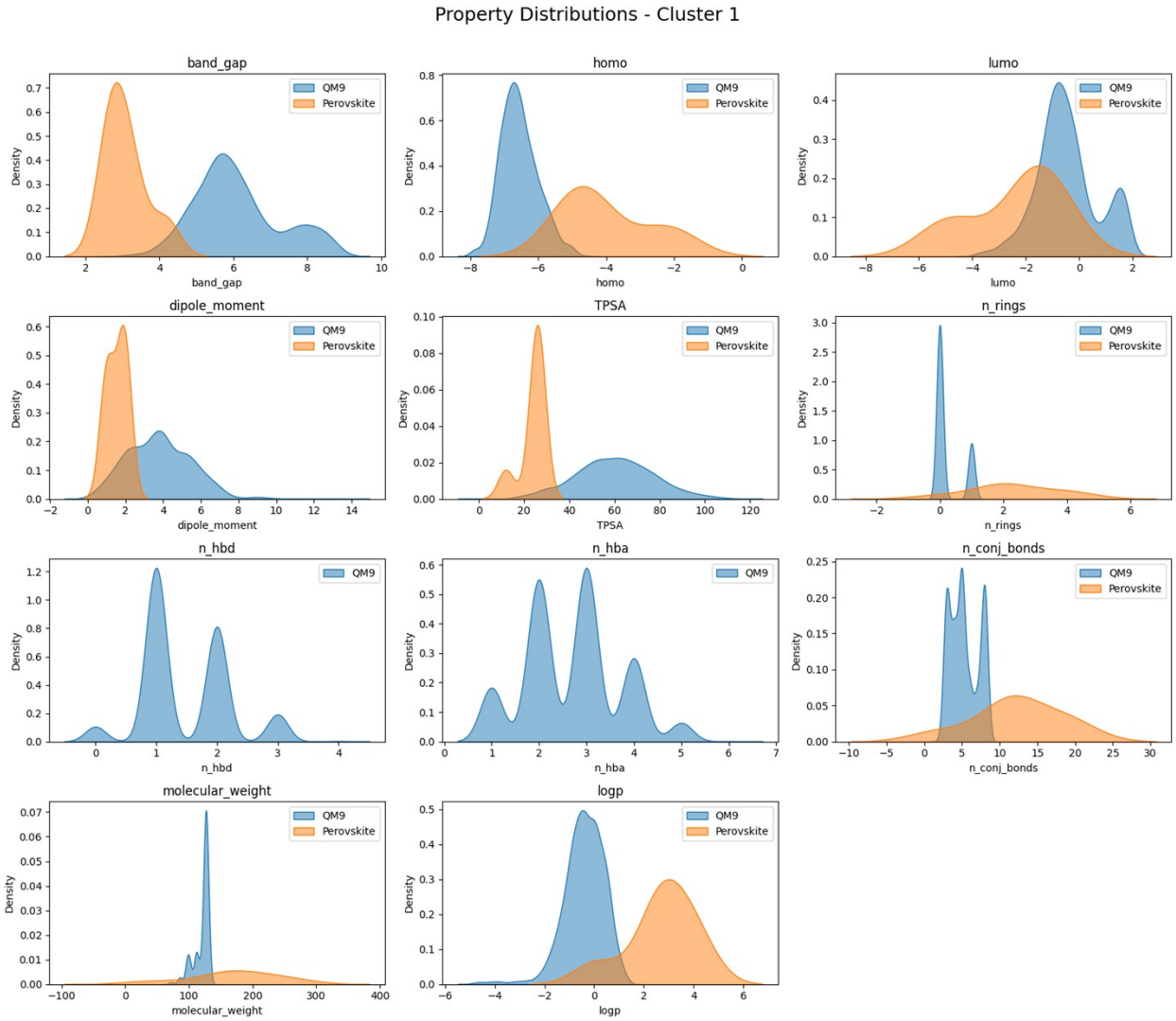

Figure S7: Property distributions for each target for cluster 1.

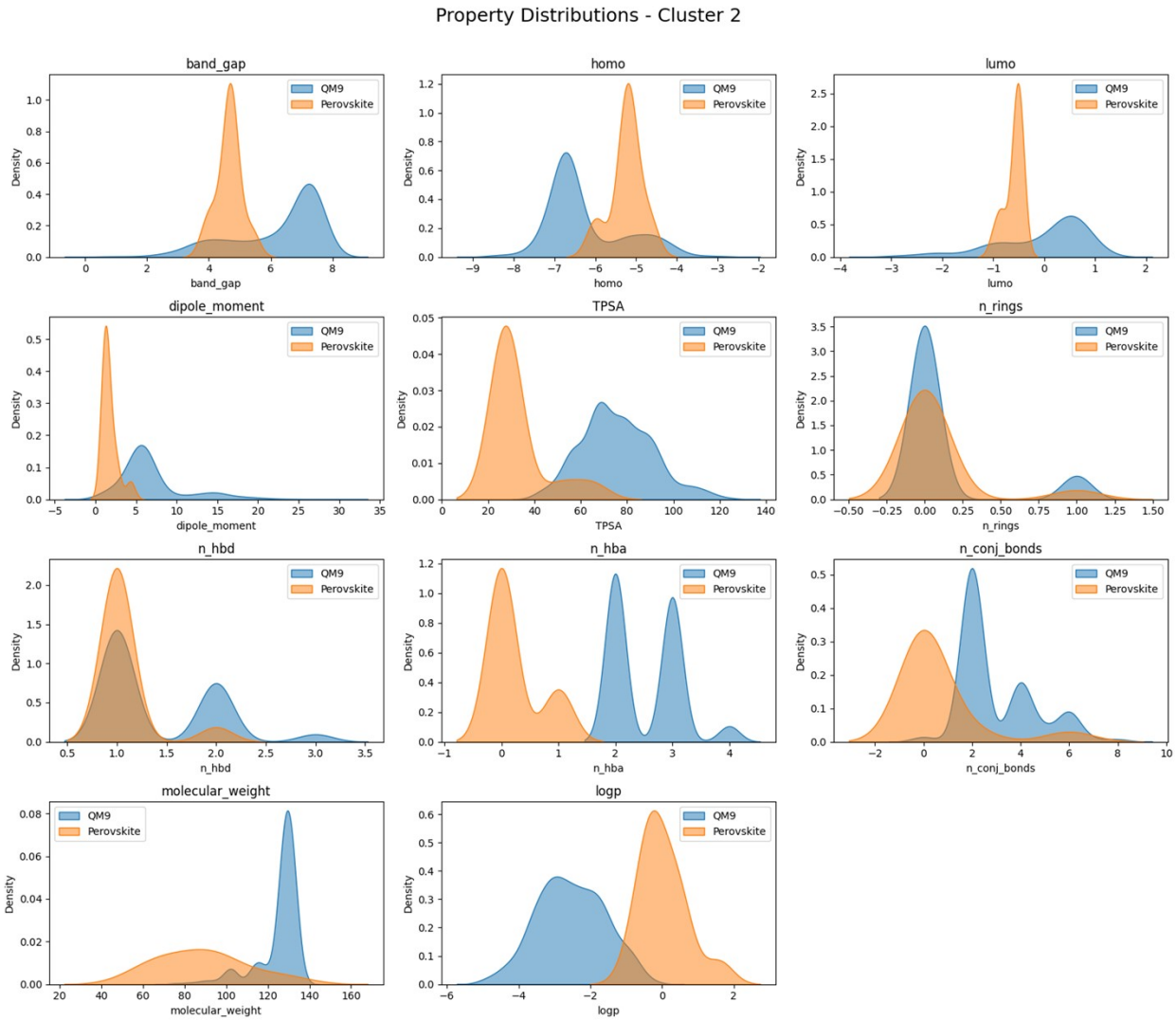

Figure S8: Property distributions for each target for cluster 2.

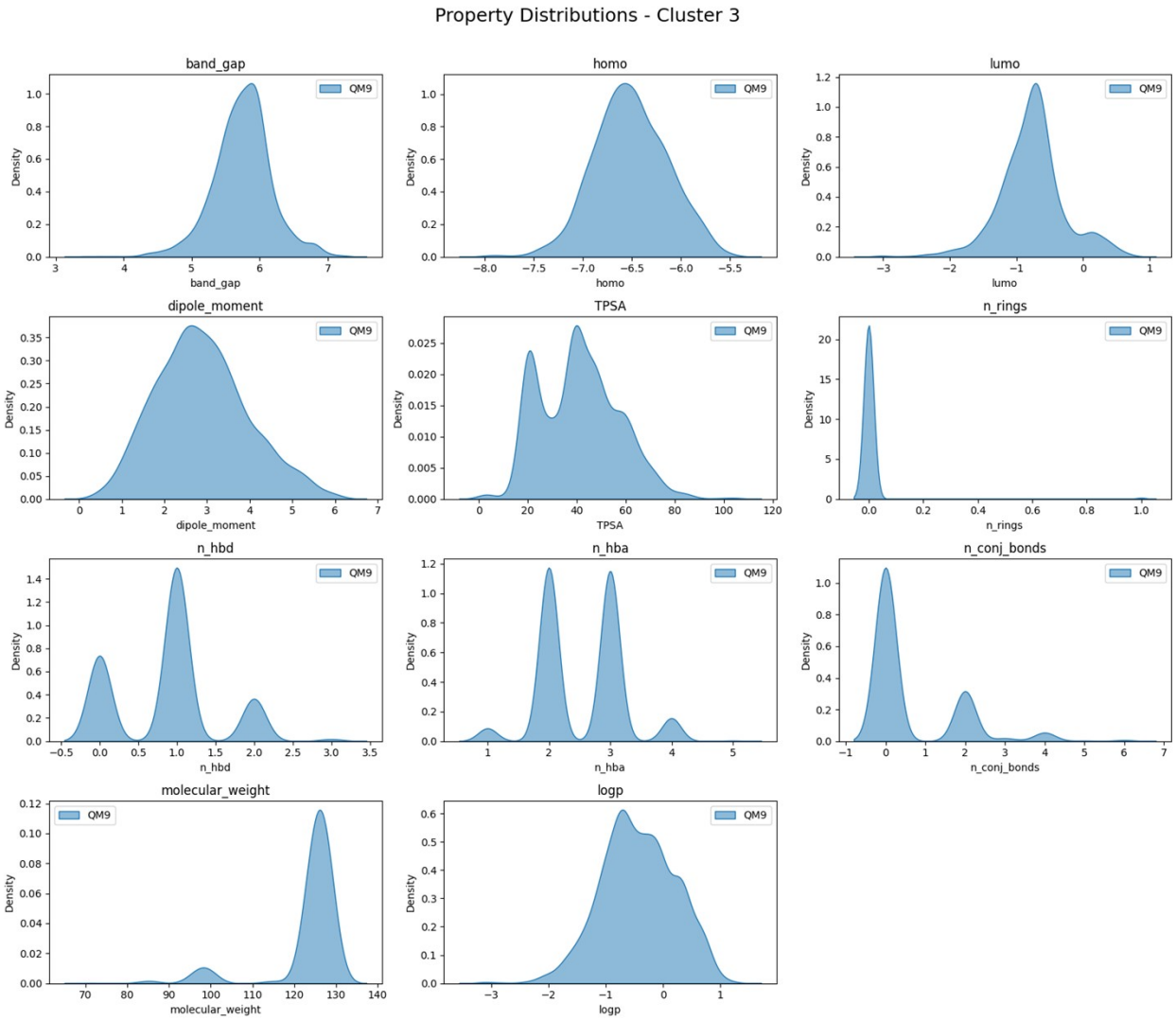

Figure S9: Property distributions for each target for cluster 3.

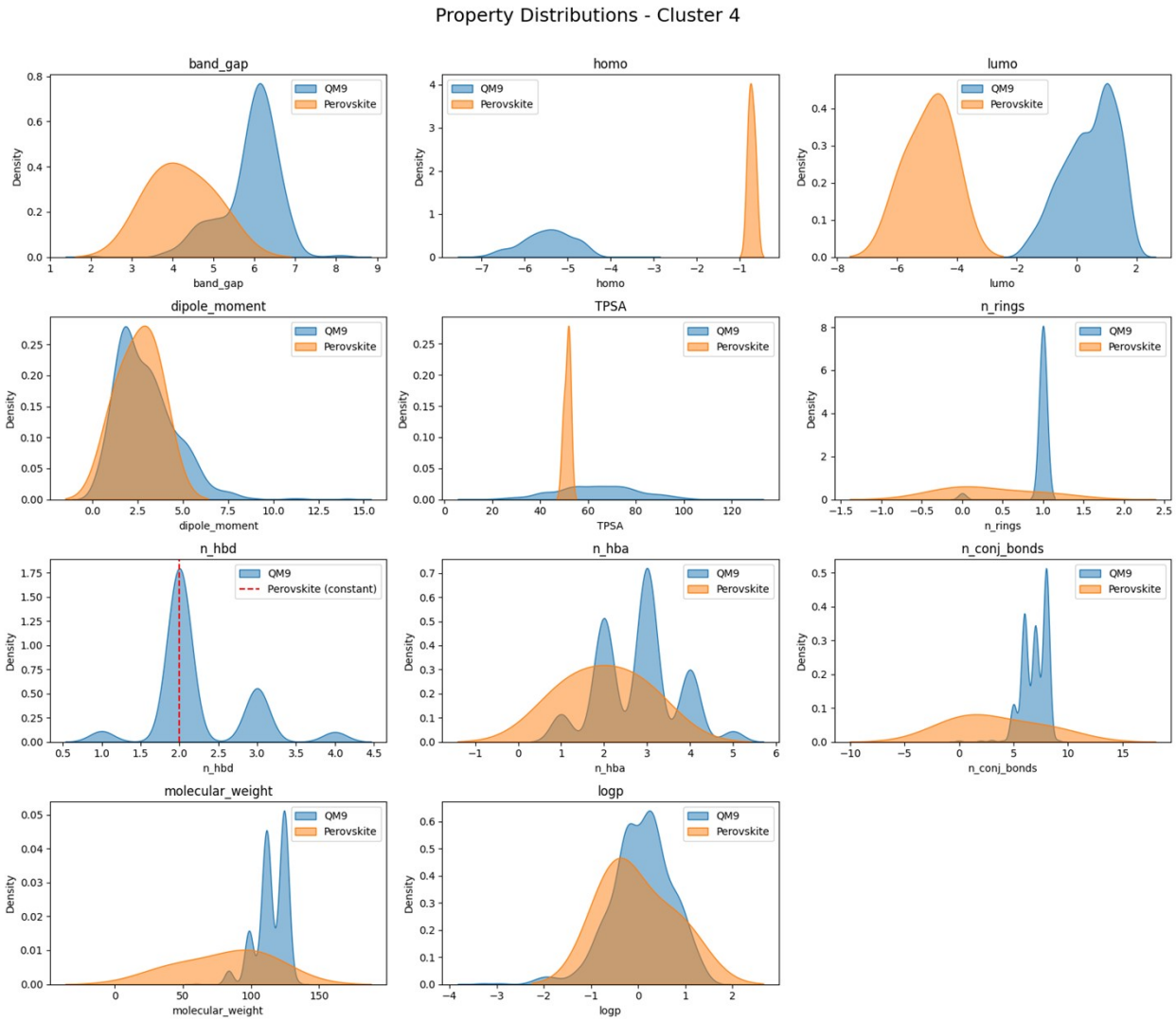

Figure S10: Property distributions for each target for cluster 4.

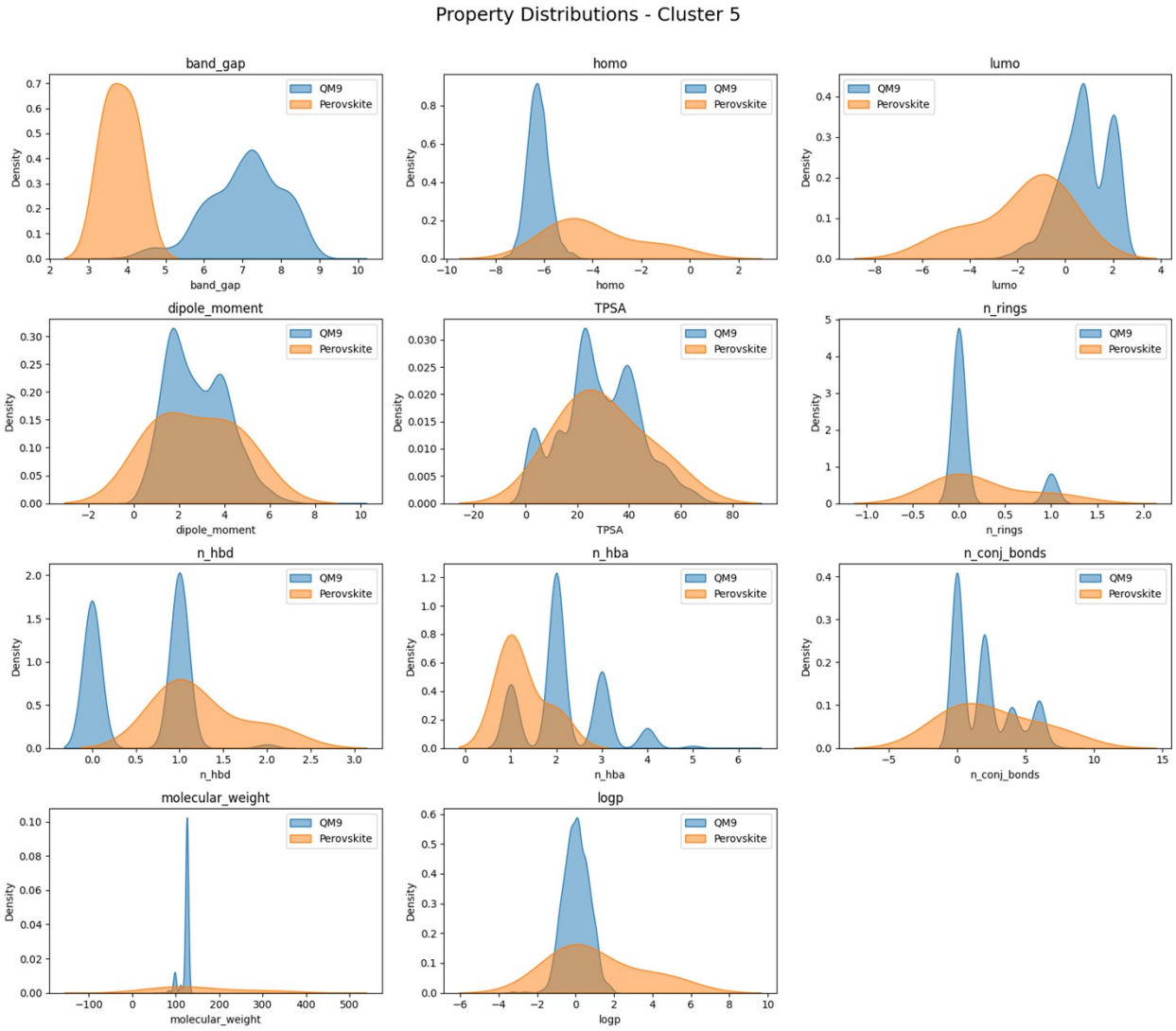

Figure S11: Property distributions for each target for cluster 5.

References:

(1) Huber, S. P.; Zoupanos, S.; Uhrin, M.; Talirz, L.; Kahle, L.; Häuselmann, R.; Gresch, D.; Müller, T.; Yakutovich, A. V.; Andersen, C. W.; Ramirez, F. F.; Adorf, C. S.; Gargiulo, F.; Kumbhar, S.; Passaro, E.; Johnston, C.; Merkys, A.; Cepellotti, A.; Mounet, N.; Marzari, N.; Kozinsky, B.; Pizzi, G. AiiDA 1.0, a Scalable Computational Infrastructure for Automated Reproducible Workflows and Data Provenance. *Sci Data* **2020**, *7* (1), 300. https://doi.org/10.1038/s41597-020-00638-4.

(2) Uhrin, M.; Huber, S. P.; Yu, J.; Marzari, N.; Pizzi, G. Workflows in AiiDA: Engineering a High-Throughput, Event-Based Engine for Robust and Modular Computational Workflows. *Computational Materials Science* **2021**, *187*, 110086. https://doi.org/10.1016/j.commatsci.2020.110086.

(3) *RDKit*. https://www.rdkit.org/ (accessed 2025-05-27).

(4) Kovács, D. P.; Moore, J. H.; Browning, N. J.; Batatia, I.; Horton, J. T.; Pu, Y.; Kapil, V.; Witt, W. C.; Magdău, I.-B.; Cole, D. J.; Csányi, G. MACE-OFF: Short-Range Transferable Machine Learning Force Fields for Organic Molecules. *J. Am. Chem. Soc.* **2025**. https://doi.org/10.1021/jacs.4c07099.

(5) Kresse, G.; Hafner, J. Norm-Conserving and Ultrasoft Pseudopotentials for First-Row and Transition Elements. *J. Phys.: Condens. Matter* **1994**, *6* (40), 8245. https://doi.org/10.1088/0953-8984/6/40/015.

(6) Kresse, G.; Joubert, D. From Ultrasoft Pseudopotentials to the Projector Augmented-Wave Method. *Phys. Rev. B* 1999, *59* (3), 1758–1775. https://doi.org/10.1103/PhysRevB.59.1758.

(7) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77* (18), 3865–3868. https://doi.org/10.1103/PhysRevLett.77.3865.

(8) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *The Journal of Chemical Physics* **2010**, *132* (15), 154104. https://doi.org/10.1063/1.3382344.

(9) Grimme, S.; Ehrlich, S.; Goerigk, L. Effect of the Damping Function in Dispersion Corrected Density Functional Theory. *Journal of Computational Chemistry* **2011**, *32* (7), 1456–1465. https://doi.org/10.1002/jcc.21759.

(10) Kresse, G.; Furthmüller, J. Efficiency of Ab-Initio Total Energy Calculations for Metals and Semiconductors Using a Plane-Wave Basis Set. *Computational Materials Science* **1996**, *6* (1), 15–50. https://doi.org/10.1016/0927-0256(96)00008-0.

(11) Kresse, G.; Furthmüller, J. Efficient Iterative Schemes for Ab Initio Total-Energy Calculations Using a Plane-Wave Basis Set. *Phys. Rev. B* **1996**, *54* (16), 11169–11186. https://doi.org/10.1103/PhysRevB.54.11169.

(12) Kresse, G.; Hafner, J. Ab Initio Molecular Dynamics for Liquid Metals. *Phys. Rev. B* **1993**, *47* (1), 558–561. https://doi.org/10.1103/PhysRevB.47.558.