Geometry-Based Neural-Network Prediction of Electron Localization Function Topology in Dense Hydrogen

Xiaoyu Wang,[*,†] Miriam Marqués,[‡] Sergio Gómez,[¶,§] Francesc Serratosa,[¶] Eva Zurek,[‖] and Julia Contreras-García[*,†]

†*Sorbonne Université, CNRS, Laboratoire de Chimie Théorique, LCT, 75005 Paris, France*

‡*CSEC, School of Physics and Astronomy, University of Edinburgh, Edinburgh, EH9 3JZ, United Kingdom*

¶*Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, 43007 Tarragona, Spain*

§*ComSCIAM, Universitat Rovira i Virgili, 43007 Tarragona, Spain*

‖*Department of Chemistry, State University of New York at Buffalo, Buffalo, NY 14260-3000, USA*

E-mail: xiaoyu.wang@sorbonne-universite.fr; julia.contreras_garcia@sorbonne-universite.fr

Abstract

We develop a machine-learning framework to predict the electron localization function (ELF) of pure, dense hydrogen directly from atomic geometry, bypassing explicit electronic-structure calculations. Trained on first-principles data spanning multiple pressure regimes in dense fluid hydrogen, the model achieves high accuracy ($R^2 > 0.99$) and faithfully reproduces the global distribution of the ELF. A combined real- and reciprocal-space analysis reveals that the residual error is dominated by smooth, long-wavelength components with correlation lengths exceeding typical H–H bonding scales,

and that the magnitude of these components increases systematically with pressure. Despite being trained exclusively on dense fluid hydrogen networks, the model transfers robustly to crystalline hydrogen configurations, preserving key features of ELF topology, including critical points and hydrogen-network connectivity. Taken together, these results suggest a viable route toward geometry-based, high-throughput evaluation of hydrogen-networking characteristics in both fluid and crystalline hydrogen.

Introduction

Since the seminal proposal by Wigner and Huntington that molecular hydrogen may transform into an atomic metallic solid under extreme compression,[1] hydrogen metallization has emerged as a central problem in dense-matter physics, where electron degeneracy, strong quantum effects, and structural complexity converge to challenge electronic-structure theory and high-pressure experiments alike.[2,3] It has long been expected to exhibit exotic properties such as high-temperature superconductivity[4] and plays a central role in models of giant-planet interiors and magnetic-field generation.[5] In the solid state, the pressure evolution of hydrogen phases has been investigated extensively through experiments[6–11] and theoretical simulations.[12–16] Despite these sustained efforts, both the metallization pressure and the microscopic mechanism driving the transition remain actively debated.

In the liquid phase, dynamic-compression and optical experiments have reported an insulator-to-conductor transition in dense hydrogen, commonly discussed in terms of a liquid–liquid transition (LLT).[17–19] However, disparate experimental diagnostics yield conflicting signatures of the transition—ranging from abrupt to smooth changes—and infer widely different transition pressures, leaving the nature of the LLT unresolved.[20–22] Complementary theoretical studies employing density functional theory (DFT) and quantum Monte Carlo (QMC) simulations have sought to clarify the LLT;[23–27] nonetheless, the long spatial and temporal correlations near the transition render these *ab initio* approaches highly sensitive to finite-size effects, simulation length, and methodological choices, preventing a definitive

characterization of the LLT.[3]

To overcome the intrinsic size and time limitations of *ab initio* molecular dynamics, machine-learning interatomic potentials trained on first-principles data have been developed to enable large-scale simulations of dense liquid hydrogen.[28,29] These models are trained on reference datasets generated from DFT and/or QMC calculations and are constructed to faithfully reproduce *ab initio* energies, forces, and stresses. As a result, they permit systematic exploration of thermodynamic observables and structural order parameters across broad regions of the phase diagram,[28,30] enable accurate determination of pressures and equations of state,[29] and provide access to free-energy landscapes, phase stability, and transport properties.[30–33] Collectively, these studies demonstrate that extending simulations to sufficiently large system sizes and long time scales is essential for disentangling genuine thermodynamic behavior from sampling and convergence effects in the liquid–liquid transition.

A notable limitation of current machine-learning frameworks is the lack of direct access to electronic bonding information. In dense liquid hydrogen, molecular character is increasingly understood as a dynamic and statistical motif rather than a collection of well-defined chemical species: ab initio studies show that H–H correlations become short-lived across the LLT and that molecular fractions evolve smoothly and remain estimator-dependent.[3,25] Consequently, short H–H distances alone do not uniquely imply stable molecular bonding in this regime. Electronic descriptors offer a complementary route to characterizing bonding beyond purely geometric criteria, most notably the electron localization function (ELF),[34] which directly quantifies electronic localization and pairing.[35] In parallel, lifetime- and cutoff-sensitive bonding diagnostics have been applied to fluid and superionic hydrogen-rich systems, where the emergence of extended hydrogen networks depends sensitively on bonding criteria and observability thresholds.[36] While bonding analysis based on electronic information is therefore essential for organizing dissociation and metallization regimes, its reliance on orbital-level quantities in *ab initio* approaches makes such descriptors computationally demanding, motivating geometry-based strategies for efficient prediction of electronic bonding

measures.

The goal of the present work is to address this bottleneck by combining ab initio molecular dynamics (AIMD) data for compressed liquid hydrogen with machine-learning predictions of the electron localization function (ELF) topology based solely on local geometric descriptors, without explicit reliance on electronic wavefunctions or orbitals. We develop a neural-network model capable of predicting the full three-dimensional ELF field with high fidelity across a wide pressure range, enabling a robust and scalable representation of the electronic localization landscape in the warm dense regime. A systematic analysis of the residual between the predicted and reference ELF fields reveals that the remaining errors are dominated by smooth, long-wavelength components, which can be efficiently captured using a band-limited Fourier representation. This separation provides a transparent characterization of pressure-dependent nonlocal contributions and offers physical insight into the emergence of long-range correlations. Beyond field-level prediction, the framework enables the identification of ELF critical points and the evaluation of hydrogen-networking descriptors,[37] providing direct access to topology-based measures of bonding organization in dense hydrogen and opening avenues for future applications to more complex hydrogen-containing systems.

Methodology

Representations

The electron localization function (ELF)[34] is typically evaluated and visualized on a real-space grid, reflecting its interpretation as a spatially resolved electronic descriptor. To predict the ELF on this grid, we represent the local atomic environment around each grid point (v) by a smooth neighbor density constructed from atomic positions and expanded in a rotation-invariant basis. For each grid point located at fractional coordinate $\mathbf{r}_v$, we define a

hydrogen-only neighbor density

$$\rho(\mathbf{r};\mathbf{r}_v) = \sum_i w(r_{vi})\delta(\mathbf{r}-\mathbf{r}_{vi}), \tag{1}$$

where $\mathbf{r}_i$ denotes the position of hydrogen atom i, $\mathbf{r}_{vi} = \mathbf{r}_i - \mathbf{r}_v$, and periodic boundary conditions are enforced. In practice, the Dirac delta function is replaced by a set of smooth basis functions, yielding a continuous and differentiable representation suitable for numerical evaluation. The weighting function

$$w(r) = \frac{1}{2}[\cos(\pi r/r_{\text{cut}}) + 1],\ r \leqslant r_{\text{cut}}, \tag{2}$$

smoothly truncates the density at a finite cutoff radius r_{cut}, ensuring continuity of both the density and its first derivative at the cutoff.

The neighbor density is expanded in a product basis of Gaussian radial functions, $R_n(r) = \exp\left[-(r-\mu_n)^2/(2\sigma^2)\right]$ with centers μ_n uniformly spanning $[0, r_{\text{cut}}]$, and real spherical harmonics, Y_{lm} up to angular momentum $l_{\max}$

$$\rho(\mathbf{r};\mathbf{r}_v) \to c_{nlm}(\mathbf{r}_v) = \sum_i R_n(r_{vi})w(r_{vi})Y_{lm}(\hat{\mathbf{r}}_{vi}). \tag{3}$$

To obtain descriptors invariant under global rotations, we compute the power spectrum of the expansion coefficients

$$P^{(l)}_{nn'}(\mathbf{r}_v) = \sum_{m=-l}^{l} c_{nlm}(\mathbf{r}_v)c_{n'lm}(\mathbf{r}_v) \tag{4}$$

and retain only the upper-triangular components $n \leqslant n'$. The final feature vector at each grid point is the concatenation of $P^{(l)}_{nn'}$ over all $l = 0, \ldots, l_{\max}$, resulting in a fixed-length,

rotation-invariant representation:

$$\mathbf{x}(\mathbf{r}_v) = \bigoplus_{l=0}^{l_{\max}} \left\{ P_{nn'}^{(l)}(\mathbf{r}_v) \right\}_{n \leqslant n'} . \tag{5}$$

This construction is closely related to the Smooth Overlap of Atomic Positions (SOAP)[38] and Behler–Parrinello symmetry-function formalisms,[39] both widely used in machine-learning studies of dense hydrogen,[28–30,32,33] but differs in being evaluated on a continuous real-space grid rather than at atomic centers.

All feature vectors are standardized using statistics computed on the training set. The target ELF values are taken directly from density-functional theory calculations and are not normalized beyond their natural $[0, 1]$ range. A grid point-wise multilayer perceptron (MLP) regressor with sigmoid output maps the local density features to the predicted ELF value at the same grid point.

Training Data Construction

The training datasets used in this work were extracted from the 3000^{th} step of each 500-atom AIMD trajectory, at which both the atomic structure and electronic degrees of freedom were fully equilibrated. Three representative volumes were selected, corresponding to pressures of 76.0, 115.1, and 138.5 GPa at 1500 K. These pressure conditions sample an ensemble of hydrogen-network topologies in which the dominant local bonding motifs evolve from molecular to atomic character. The resulting distribution of configurations, shown in Fig. 1A, spans a regime where progressive electronic delocalization and fluid metallization have been extensively discussed.[3] Even at lower pressures, intermolecular interactions already perturb the idealized molecular limit, reducing and broadening the bond-centered ELF maximum and introducing intermediate ELF values.[40] Importantly, the AIMD configurations naturally span the ELF regimes that govern the networking analysis: the networking value is determined by connectivity transitions at intermediate ELF isovalues rather than by regions where ELF

≈ 1.[37,41]

The machine-learning training set was constructed by combining the ELFs calculated for three independent MD snapshots, obtained at the $P - T$ conditions noted, and calculated on a 192^3 real-space grid. In each case, 50,000 grid points, and therefore local atomic environments, were randomly chosen using a stratified scheme that is approximately uniform in ELF value, yielding a total training set of 150,000 training points. Structural descriptors were generated within a radial cutoff of 3.0 Å, using 10 radial basis functions and spherical harmonics up to $l_{\max} = 2$. All features were stored in half precision (float16) to reduce memory footprint. Sampling was performed in batches of 1024 with a fixed random seed to ensure reproducibility.

It is worth noting that all reference calculations in this study are performed using the PBE exchange–correlation functional. Consequently, the predicted ELF inherits the known limitations of this approximation, and the achievable model accuracy is fundamentally constrained by the fidelity of the underlying electronic structure description. While different functionals may shift the pressure at which the molecular-to-network transition occurs, the associated evolution of ELF topology and connectivity is expected to remain qualitatively robust.

Model Architecture and Training Protocol

We employ a multilayer perceptron (MLP) regressor[42] consisting of two hidden layers of width 128 with smooth rectified linear unit (SiLU) activations, implemented in PyTorch.[43] The model maps the descriptor vector $\mathbf{X} \in \mathbb{R}^F$ to a scalar target via a fully connected architecture. The network is trained for 80 epochs using the AdamW optimizer[44] with learning rate 3×10^{-4} and weight decay 10^{-6}. Optimization minimizes the Huber loss[45] with parameter $\beta = 0.03$, which provides a quadratic penalty for small residuals and a linear penalty for large residuals, improving robustness to outliers while retaining sensitivity near the optimum.

Training is performed with a batch size of 4096. All input features are standardized to zero mean and unit variance using statistics computed on the training set only, thereby preventing information leakage into validation data. Model parameters are initialized with a fixed random seed to ensure reproducibility. Hyperparameters—including network width, depth, cutoff radius r_{cut}, number of radial basis functions n_{radial}, maximum angular momentum l_{max}, and the Huber loss parameter—are systematically optimized via controlled one-factor-at-a-time scans, as detailed in Section S2.

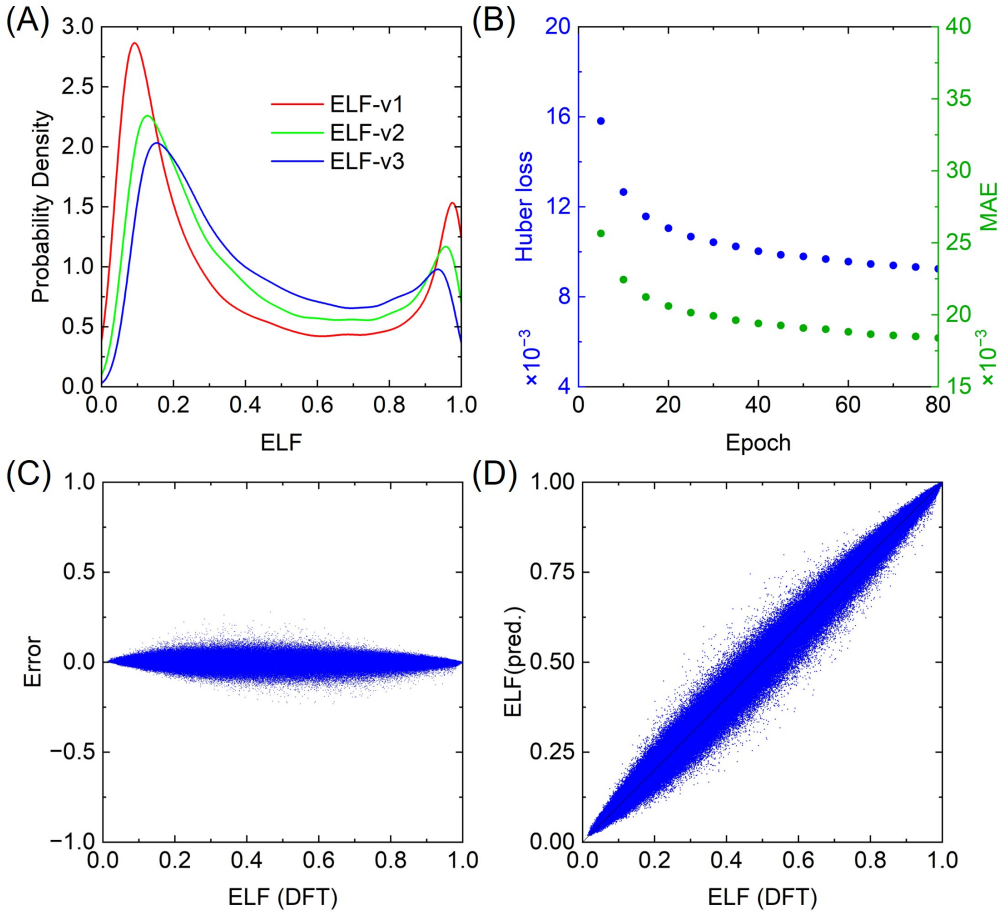


Figure 1: Data distribution, training convergence, and predictive performance of the ELF model. (A) Probability density distribution of ELF values across the dataset, illustrating the broad sampling of the target space. (B) Training Huber loss and validation mean absolute error (MAE) as a function of training epoch. (C) Prediction error as a function of the reference DFT ELF value. (D) Predicted versus reference DFT ELF values, with the solid line indicating the identity.

Results and discussion

Performance

The model was trained for 80 epochs, after which both the training Huber loss and the validation mean absolute error (MAE) reach clear plateaus, indicating stable convergence without evidence of overfitting (Figure 1B). The concurrent saturation of these metrics suggests that the representational capacity of the local-descriptor model is fully exploited within this training window.

The validation set was generated using the same sampling protocol as the training data, however in this case 150,000 local atomic environments were extracted from the dataset at each of the three pressure conditions. Figures 1C and 1D summarize the predictive performance over the training set. The error distribution as a function of the reference ELF (Figure 1C) is narrowly centered around zero across the full ELF range, with no discernible systematic bias at either low or high localization values. Correspondingly, the predicted ELF values closely follow the identity line (Figure 1D), yielding excellent agreement with the DFT reference statistics (MAE = 0.0190 , root mean squared error (RMSE) = 0.0271, and coefficient of determination (R^2) = 0.992). The predicted and reference distributions are nearly indistinguishable, with identical means (0.426) and very similar variances, confirming that the model accurately reproduces both the central tendency and overall spread of ELF values. The stability of the model was further evaluated by repeating the entire workflow (including dataset construction, training, and validation) 50 times with different random seeds. Although trained on the 3000th snapshot of each AIMD run, the model achieves comparable accuracy on the 1000th and 2000th snapshots used as an independent validation set (Table S2). The results demonstrate high reproducibility, with negligible variation in all reported metrics. Further details are provided in Section S3 of the Supporting Information.

Despite the high overall fidelity, the residual field is not purely stochastic. As shown in Fig. 1C, the prediction error varies across the ELF range, with larger absolute deviations

observed at intermediate ELF values and a correspondingly broader spread around the identity line in Fig. 1D. This region corresponds to electronic environments with ELF values close to that of a homogeneous electron gas (HEG, ELF ≈ 0.5), where electron localization is weak and bonding characteristics are less clearly defined. In this regime, relatively small geometric or electronic variations can lead to noticeable changes in ELF, making prediction more challenging. In contrast, highly localized (ELF $\to 1$) and strongly depleted (ELF $\to 0$) regions exhibit more distinct signatures and are predicted with higher accuracy. The enhanced deviations at intermediate ELF values therefore motivate a more detailed analysis of the structured residuals, which is presented in the following section.

Origin of the Residual

To clarify the physical origin of the residual field not captured by the local descriptor model, we examine its real-space structure using a representative two-dimensional slice from the 76.0 GPa configuration (Fig. 2A). The corresponding residual field (Fig. 2B) exhibits a weak, smoothly varying modulation extending over a substantial fraction of the unit cell. The residual magnitude is uniformly small, with most values confined within 0.05 and only rare, spatially diffuse regions approaching larger deviations. These features span several tenths of the unit cell, corresponding to correlation lengths of a few angstroms, well above typical bond lengths, indicating a low-amplitude, long-wavelength contribution rather than missing short-range or chemically specific effects.

This picture is reinforced by reciprocal-space analysis. The two-dimensional Fourier spectrum of the residual field (Fig. 2C) is strongly Γ-centered and nearly isotropic, with spectral weight concentrated at small wavevectors and smoothly decaying toward the Brillouin-zone boundary. No anisotropy, lattice-locked features, or enhancement at large $|\mathbf{k}|$ are observed. Consistently, the radially averaged Fourier amplitude (Fig. 2D) decreases monotonically with increasing $|\mathbf{k}|$, without secondary peaks or characteristic length scales. These observations demonstrate that the residual is dominated by coherent long-wavelength contributions not

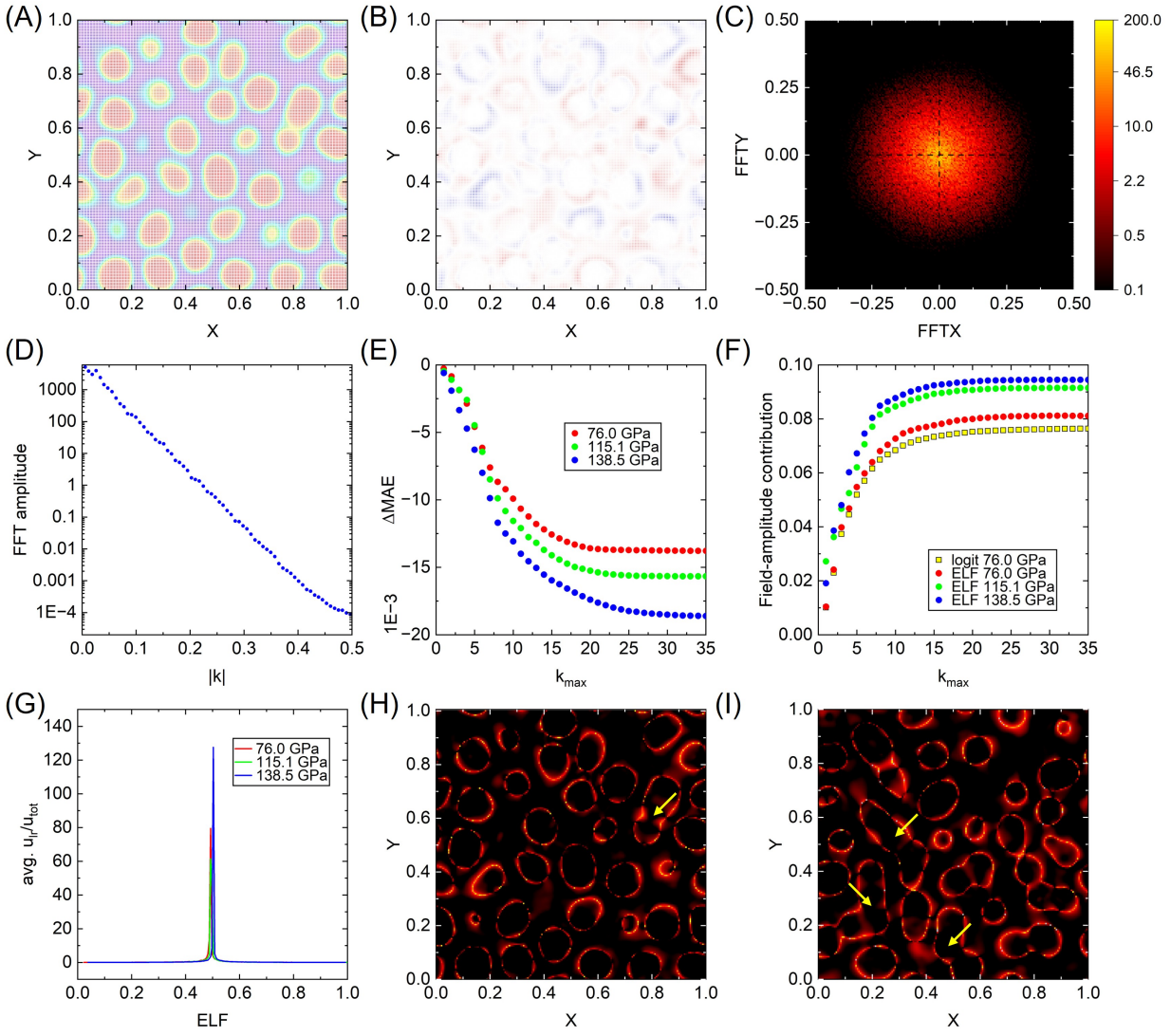


Figure 2: ELF and residual-field analysis for a representative two-dimensional slice of the 76.0 GPa structure. (A) DFT-computed ELF in real space (color scale from 0 (blue) to 1 (red); fractional coordinates relative to 11 Å). (B) Real-space residual between predicted and DFT ELF (color scale from −0.2 (blue) through 0 (white) to +0.2 (red)). (C) Two-dimensional Fourier transform of the residual field. (D) Radially averaged Fourier amplitude as a function of wave vector k. (E) Change in mean absolute error (ΔMAE) versus Fourier cutoff $k_{\max}$ for 76.0, 115.1, and 138.5 GPa. (F) Fractional magnitude of the Fourier correction versus $k_{\max}$, evaluated in logit space at 76.0 GPa (yellow squares) and in ELF space at 76.0, 115.1, and 138.5 GPa (circles). (G) Distribution of the Fourier-correction contribution ($u_{\mathrm{lr}}/u_{\mathrm{tot}}$) as a function of ELF. (H) Real-space distribution of the Fourier-correction contribution at 76.0 and (I) 138.5 GPa (logarithmic color scale from 0.1 (black) to 25(yellow)).

captured by strictly local descriptors, establishing a clear separation between local and collective components of the ELF.

Motivated by this clear separation of length scales, the residual can be systematically reduced by introducing a HEG-like background as a long-range correction. The correction is naturally formulated in the logit representation of the ELF, which provides an unbounded scalar field suitable for additive decomposition. The real-space logit associated with the locally predicted ELF, y_{loc}, is defined as

$$u_{\mathrm{loc}}(\mathbf{r}) = \mathrm{logit}(y_{\mathrm{loc}}(\mathbf{r})) = \ln\left[\frac{y_{\mathrm{loc}}(\mathbf{r})}{1 - y_{\mathrm{loc}}(\mathbf{r})}\right], \tag{6}$$

where $\mathbf{r}$ denotes the fractional coordinate within the unit cell. Within this representation, long-range contributions, $u_{\mathrm{lr}}(\mathbf{r})$, are expressed as a smooth, band-limited Fourier expansion as,

$$u_{\mathrm{lr}}(\mathbf{r}) = \sum_{|\mathbf{k}|<k_{\mathrm{cut}}} \left[a_{\mathbf{k}} \cos(2\pi\mathbf{k}\cdot\mathbf{r}) + b_{\mathbf{k}} \sin(2\pi\mathbf{k}\cdot\mathbf{r})\right], \tag{7}$$

where k_{cut} sets the maximum spatial frequency retained in the long-range field, and the coefficients $a_{\mathbf{k}}$ and $b_{\mathbf{k}}$ are determined by a least-squares fit to the residual. By construction, this expansion captures only smooth, collective variations while excluding short-range, atom-centered features already described by u_{loc}.

The total logit field is obtained through the additive decomposition

$$u_{\mathrm{tot}}(\mathbf{r}) = u_{\mathrm{loc}}(\mathbf{r}) + u_{\mathrm{lr}}(\mathbf{r}), \tag{8}$$

and the corrected ELF is recovered by mapping back to the physical interval $[0, 1]$ using the inverse logit transformation:

$$y_{\mathrm{pred}}(\mathbf{r}) = \frac{1}{1 + \exp[-u_{\mathrm{tot}}(\mathbf{r})]}. \tag{9}$$

We emphasize that this long-range correction is not introduced as a practical solution for

improving transferable ELF predictions. Rather, it serves as an interpretive tool that exposes the separation between short-range contributions captured by strictly local geometric descriptors and residual long-wavelength components that reflect collective, nonlocal electronic effects. By isolating these smooth contributions in a controlled manner, the analysis provides physical insight into the nature and pressure evolution of nonlocal correlations, rather than constituting an additional predictive model.

Taking the two-dimensional slice discussed above as a representative example, the analysis is performed on a 192×192 real-space grid extracted from the 76.0 GPa structure. For this slice, the purely local-descriptor baseline yields an MAE of 0.0184 (RMSE $\approx$ 0.0283). Figure 2E shows the evolution of the MAE reduction, ΔMAE, as a function of the maximum Fourier cutoff $k_{\max}$ used to construct the long-range correction in logit space. As $k_{\max}$ increases, ΔMAE decreases monotonically and converges rapidly, reaching a clear plateau at $k_{\max} \approx 20$. Beyond this cutoff, no further improvement is observed, and the MAE reduction saturates at $\Delta\text{MAE} \approx -0.0137$ ($\Delta\text{RMSE} \approx -0.022$), corresponding to a remaining error of MAE $\approx$ 0.0047 and RMSE $\approx$ 0.0063. The progressive improvement of the residual maps with increasing $k_{\max}$ is shown in Figure S3. This convergence behavior demonstrates that the residual error of the local-descriptor model is almost entirely accounted for by a band-limited, long-wavelength field.

The quantitative contribution of the long-range correction is summarized in Fig. 2F. Despite its pronounced impact on prediction accuracy, the long-range component contributes only a small fraction of the total field amplitude: approximately 7.6% of the total logit field u_{tot} and 8.1% of the corrected ELF field y_{pred} at 76.0 GPa. This apparent disparity highlights an important physical point: although small in magnitude, the coherent low-k structure of the long-range component enables it to correct systematic errors that are inaccessible to strictly local descriptors. The residual therefore represents a genuinely nonlocal contribution—minor in amplitude, yet essential for quantitative accuracy.

A clear pressure dependence is observed in the magnitude and spatial extent of the long-

wavelength residual. The same analysis, performed on analogous two-dimensional slices at 115.1 and 138.5 GPa, yields baseline MAEs of 0.0192 and 0.0197, respectively, indicating progressively stronger long-range residuals at higher pressure. Accordingly, convergence of the Fourier correction requires progressively higher cutoffs, with $k_{\max} \approx 24$ at 115.1 GPa and $k_{\max} \approx 27$ at 138.5 GPa, consistent with a more plane-wave-like character of the residual field (Figure 2E). The fractional contribution of the correction in ELF space also increases to approximately 9.1% and 9.4%, respectively (Figure 2F).

We next examine the ELF- and real-space structure of this correction. As shown in Figure 2G, the Fourier correction is strongly concentrated around ELF$\approx$0.5, indicating that the long-wavelength contribution primarily resides in electronically intermediate regions rather than at fully localized maxima or minima. This pressure-driven evolution is also evident in real space. Figure 2H shows the spatial distribution of the Fourier-correction contribution (corresponding to Figure 2A at 76.0 GPa). The red contours typically enclose a single ELF localization center, consistent with a regime dominated by localized bonds or lone-pair–like features at this pressure (and 1500 K), with polymerized hydrogen motifs occurring only rarely. Upon increasing the pressure to 138.5 GPa, an extended pattern of polymerization emerges across large regions of the cell: the contours connect neighboring ELF maxima, forming continuous networks that span multiple localization centers (Figure 2I). These contours represent the spatial weighting of the retained Fourier components, thereby highlighting electronic features that require intrinsically delocalized descriptors. Consistent with this interpretation, exploratory convolutional models[46] indicate that ELF prediction is dominated by short-range contributions (Section S1), and that the remaining long-wavelength residual arises from intrinsic locality limitations rather than model inadequacy. Taken together, these observations reveal a clear mechanism for pressure-induced metallization in hydrogen networks, whereby progressive electronic delocalization manifests as the growth, compression, and interconnection of long-range correlation structures in real space.

ELF Networking Validation

The training data employed in this work are obtained from AIMD simulations of dense hydrogen, a setting that naturally enables detailed analysis of bonding evolution and phase transitions under high-pressure and high-temperature conditions, and within which both the model accuracy and the physical origin of the residual have been established. A further, more challenging step is to examine the predictive power of the model with respect to ELF topology, including the extraction of critical points in ground-state structures to quantify hydrogen-framework networking, defined as the highest ELF isovalue at which a continuous, crystal-spanning network of electronic localization—mediated by ELF saddle points—is formed.[37] This capability would enable rapid estimates of the superconducting critical temperature (T_c), as implemented in the TcEstime framework.[41] Such a pipeline is particularly attractive as a filtering stage in crystal-structure-prediction (CSP) workflows or random-structure searches targeting superconducting hydrides.[47,48] For these use cases, performance must be evaluated on randomly generated structures to assess transferability outside the training domain.

To this end, we generated cubic and hexagonal hydrogen lattices using the RandSpg code.[49] A total of 119 cubic and 230 hexagonal structures were constructed and fully optimized at 100 GPa. Only local geometric descriptors were employed in the present model, as no transferable global long-range correction field was trained across distinct geometries. The fitting performance for both crystal families is summarized in Fig. 3A and B, where the R^2 of the pointwise ELF prediction is analyzed as a function of the networking value. Two complementary metrics are reported: the identity R^2 measures agreement with the identity line, and the correlated R^2 which measures linear correlation between predicted and DFT ELF values independent of deviations from the identity line. A low identity R^2 combined with a high correlated R^2 therefore indicates good linearity of the prediction but a systematic deviation from the identity. Both cubic and hexagonal structures exhibit good overall agreement, with average identity R^2 values of 0.96 and 0.94, respectively. A systematic deviation

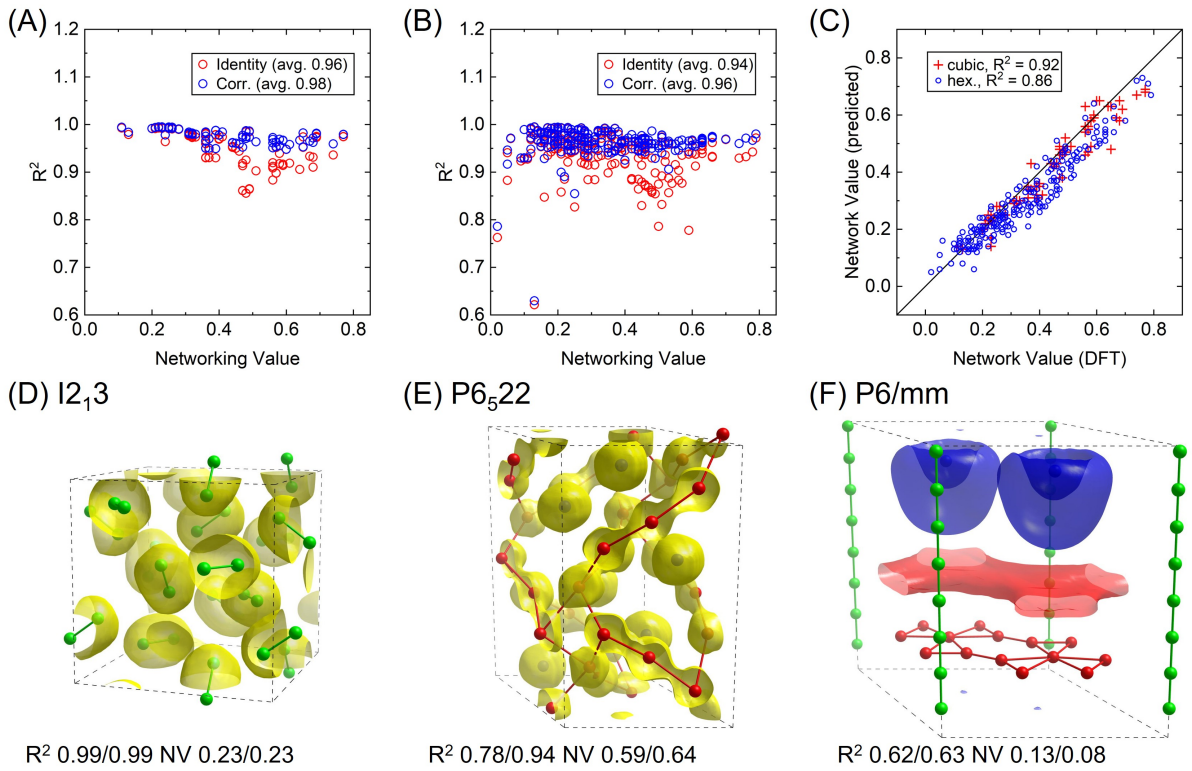


Figure 3: Performance of the local-descriptor ML model for predicting ELF-derived hydrogen networking values. (A,B) Coefficient of determination (R^2) of the pointwise ELF prediction error for individual structures, shown as a function of the networking value for (A) cubic and (B) hexagonal structures optimized at 100 GPa. (C) ML-predicted versus DFT networking values for cubic (red) and hexagonal (blue) systems; the solid line denotes perfect agreement. (D) Cubic $I2_13$ structure with isolated H_2 units (ELF isovalue 0.85). (E) Hexagonal $P6_522$ structure featuring a three-dimensional hydrogen framework and isolated H^- species (ELF isovalue 0.65). (F) Hexagonal $P6/mm$ structure with one-dimensional hydrogen chains and extended interstitial regions, shown as ΔELF at isovalue 0.25. Green indicates H_2 molecules or polymeric chains, red the three-dimensional hydrogen framework, and blue H^- species. The annotations in panels (D–F) are reported as "R^2 (identity/correlated), ϕ (ML/DFT)", where the identity R^2 measures agreement with the identity line, and the correlated R^2 measures linear correlation independent of deviations from the identity. ϕ denotes the networking value predicted by ML and computed from DFT, respectively.

between identity and correlated R^2 is observed near intermediate networking values (~ 0.5), consistent with the increased difficulty of accurately capturing weakly localized electronic environments, as discussed above.

The direct prediction of networking values remains robust despite the increased difficulty at intermediate ELF values, as shown in Fig. 3C, where the overall R^2 between predicted and DFT networking values is 0.92 for cubic structures and 0.86 for hexagonal structures. A stability analysis was performed for ELF prediction and networking value (NW) evaluation on randomly generated structures. For each structure, 50 independent model realizations were used. The results confirm that the run-to-run variability remains small, while the variation across different structures reflects intrinsic differences in structural complexity and prediction difficulty. Detailed statistics are provided in Section S3 and in the accompanying supplementary table (xlsx file). For molecular hydrogen structures, the networking value is small because electrons are strongly localized within intramolecular H–H bond pairs, resulting in weak electronic connectivity between neighboring molecules. A large subset of the cubic structures considered here falls into this regime and exhibits minimal systematic prediction errors (Fig. 3A). A representative example is the cubic $I2_13$ structure shown in Fig. 3D, which consists exclusively of isolated H_2 molecules. In this case, the networking value is 0.23 in both the DFT reference and ML prediction, and both identity and correlated R^2 values reach 0.99. These results indicate that the ELF in molecular phases is accurately captured by purely local geometric representations.

A more challenging regime is illustrated by the $P6_522$ structure (Fig. 3E). In this case, a clear separation emerges between the identity and correlated metrics: the predicted ELF exhibits strong linear correlation with the DFT reference (correlated $R^2 = 0.94$) but shows a noticeable deviation from the identity relation (identity $R^2 = 0.78$), indicating a systematic offset in the absolute ELF values. This structure features a three-dimensional, delocalized ELF network coexisting with isolated hydrogen units, giving rise to extended electronic connectivity beyond strictly local environments. Such delocalized ELF topologies are inher-

ently more difficult to reproduce with purely local geometric descriptors, leading to reduced agreement with the identity line despite preserved linearity. Nevertheless, the predicted networking value (0.64) remains in close agreement with the DFT value (0.59), demonstrating that the essential topological connectivity of the hydrogen framework is correctly captured even when absolute ELF values deviate systematically.

A quick estimation of the superconducting critical temperature can be obtained within the TcEstime framework. For the present pure hydrogen systems, the hydrogen fraction and the hydrogen-resolved density of states at the Fermi level are trivially unity, such that the descriptor reduces directly to the networking value (ϕ). Using the empirical relation proposed by Belli *et al.*,[37] we obtain estimated critical temperatures spanning from $\sim$ 0–400 K across representative structures (Table 1). The close agreement between ML-predicted and DFT-derived values reflects the high fidelity of the model at the level of the networking descriptor, demonstrating that the physically relevant trends for superconductivity screening are preserved within this framework.

Table 1: Estimated superconducting critical temperatures T_c obtained from the networking value (ϕ) using the TcEstime relation $T_c = 750\,\phi - 85$ K. Negative values are interpreted as non-superconducting ($T_c \approx 0$ K).

Structure	ϕ (DFT)	ϕ (ML)	T_c (DFT) [K]	T_c (ML) [K]
$I2_13$	0.23	0.23	87	87
$P6_522$	0.59	0.64	358	395
$P6/mm$	0.13	0.08	13	-25

The most severe failure mode is observed exclusively in a subset of hexagonal structures. Lower symmetry enables more complex hydrogen topologies, as exemplified in Figure 3F. In this structure, one-dimensional hydrogen chains extend along the z direction, while the interchain regions are occupied by alternating layers of isolated atomic hydrogen two-dimensional hydrogen networks. Analysis of the ELF residual reveals the origin of the error: the large interstitial regions between the hydrogen atoms and the 2D network are essentially invisible to the local descriptors, resulting in a negative ELF deviation around the H^- sites and a compensating positive deviation in the interstitial region. Consequently, both identity and

correlated R^2 values deteriorate to 0.62–0.63. This case highlights a genuine limitation of the present model, arising from the absence of large, electronically active interstitial regions in the dense-hydrogen training data. In practice, however, such structures are likely to be eliminated by complementary CSP filters (e.g., energetic stability) and are therefore unlikely to survive in energy-driven structure searches.[50]

Overall, these tests demonstrate that the present ML model provides reliable predictions of hydrogen networking values across a broad range of randomly generated structures. Although global background contributions can limit pointwise ELF accuracy in specific cases, the preserved linearity ensures that derived networking values remain quantitatively meaningful. Beyond static structure screening, this framework could be naturally coupled with machine-learning interatomic potentials to enable large-scale molecular-dynamics simulations of dense hydrogen, in which bonding topology and hydrogen-network connectivity can be analyzed without explicit *ab initio* calculations. Beyond elemental hydrogen, it is reasonable to anticipate that this approach may extend to metal-doped hydrogen-rich compounds, where the dominant bonding topology and connectivity are primarily governed by the hydrogen sublattice, while the metal species play a secondary structural or charge-balancing role. Together, these capabilities suggest a viable route toward alleviating a major computational bottleneck in the exploration of hydrogen-rich and multicomponent chemical spaces.[51]

Conclusions

We have demonstrated that a machine-learning model based solely on local geometric descriptors can predict the electron localization function (ELF) of dense hydrogen-rich systems with high accuracy across multiple pressure regimes, achieving $R^2 > 0.99$ while faithfully reproducing the global ELF distribution. A systematic analysis of the residual reveals that the remaining error is not stochastic, but arises from smooth, long-wavelength components with correlation lengths exceeding typical bond distances, whose magnitude and spatial

extent increase with pressure. These nonlocal contributions are efficiently captured by a band-limited Fourier correction formulated in the logit representation, despite accounting for only a small fraction of the total field amplitude. Importantly, ELF-derived topological descriptors—specifically hydrogen-networking values—remain robust in the presence of structured residuals, enabling reliable characterization of hydrogen bonding topology across a wide range of configurations. While the present study focuses on elemental hydrogen, these observations suggest that similar approaches may be applicable to hydrogen-rich compounds in which the dominant bonding connectivity is governed by the hydrogen sublattice, motivating future investigations in more complex chemical environments.

A key advantage of the present framework is that it completely bypasses the explicit calculation of Kohn–Sham orbitals or kinetic-energy densities, which are traditionally required for ELF evaluation. Instead, ELF is inferred directly from atomic geometry, yielding orders-of-magnitude reductions in computational cost relative to conventional first-principles workflows. Earlier attempts to approximate ELF using density-only formulations have been shown to suffer from limited accuracy and sensitivity to the choice of reference frame, restricting their practical applicability.[52] By contrast, the present approach achieves high fidelity without recourse to wavefunction-level information, making it well suited for high-throughput and exploratory studies where direct electronic-structure calculations would be computationally prohibitive.

Computational method

Ab initio calculations

AIMD data were obtained from simulations reported previously,[35] performed in cubic supercells containing 500 hydrogen atoms. Electronic-structure calculations were carried out within density functional theory using the Perdew–Burke–Ernzerhof (PBE) exchange–correlation functional,[53] together with the projector–augmented-wave (PAW) method,[54] as implemented

in the Vienna *ab initio* Simulation Package (VASP).[55] A plane-wave kinetic-energy cutoff of 700 eV was employed. Brillouin-zone sampling was restricted to the Baldereschi mean-value point,[56] which has been shown to provide accuracy comparable to that of a $4 \times 4 \times 4$ Monkhorst–Pack grid for liquid hydrogen at similar system sizes. All AIMD simulations were performed in the canonical (NVT) ensemble using a time step of 0.5 fs and a Nosé–Hoover thermostat[57,58] for temperature control. For each thermodynamic condition, the system was equilibrated for 1 ps, followed by a 1.5 ps production run.

Crystalline hydrogen structures were randomly generated using the RandSpg code for all cubic and hexagonal space groups, with 72 atoms per unit cell.[49] Geometry optimizations were performed at 100 GPa using the PBE functional within VASP. Electrons were represented using PAW pseudopotentials with a plane-wave cutoff energy of 600 eV. Reciprocal space was sampled using a Γ-centered k-point mesh with a maximum spacing of 0.15 Å^{-1}. Topological analysis of ELF isosurfaces was carried out using the Critic2 package,[59] and hydrogen-framework networking values were evaluated using the TcEstime framework.[41]

Data Availability

The code used in this study is publicly available at `https://github.com/July13210914/ELF-Prediction`. All scripts required to reproduce the training and evaluation procedures are included in the repository. The data that support the findings of this study are available from the online repository https://www.lct.jussieu.fr/pagesperso/contrera/databaseELFprediction/

Acknowledgments

This work was supported by the Agence Nationale de la Recherche (ANR) under Grant No. ANR-22-CE50-0014 and by the ECOS-Sud program under Projects C17E09 and C21E06/ECOS210019.

Computational resources were provided by GENCI under Projects No. A0160915069, A0160815101, and A0190915069, UKCP Archer at EPCC (EPSRC Grant No. EP/P022790/1)

and by the Center for Computational Research at SUNY Buffalo (http://hdl.handle.net/10477/79221).

S.G. acknowledges support from MICIN PID2021-128005NB-C21 and RED2022-134890-T, Generalitat de Catalunya 2021SGR-633, and Universitat Rovira i Virgili 2025INTER-03 and 2023PFR-URV-00633. M.M. acknowledges support from the ERC fellowship "Hecate". E.Z. acknowledges the U.S Department of Energy, Office of Science, Fusion Energy Sciences funding the award entitled High Energy Density Quantum Matter, under Award No. DE-SC0020340. F.S. acknowledges support from the project PID2022-138327OB-I00, financed by the Ministerio de Ciencia e Innovación (MCIN)/Agencia Estatal de Investigación (AEI)/10.13039/501100011033/FEDER, UE.

Conflict of interests

The authors declare no conflict of interest.

References

(1) Wigner, E.; Huntington, H. On the possibility of a metallic modification of hydrogen. *J. Chem. Phys.* **1935**, *3*, 764–770.

(2) Gregoryanz, E.; Ji, C.; Dalladay-Simpson, P.; Li, B.; Howie, R. T.; Mao, H.-K. Everything you always wanted to know about metallic hydrogen but were afraid to ask. *Matter Radiat. Extrem.* **2020**, *5*.

(3) Bonitz, M. et al. Toward first principles-based simulations of dense hydrogen. *Phys. Plasmas* **2024**, *31*.

(4) Ashcroft, N. W. Metallic hydrogen: a high-temperature superconductor? *Phys. Rev. Lett.* **1968**, *21*, 1748.

(5) Guillot, T. The interiors of giant planets: Models and outstanding questions. *Annu. Rev. Earth Planet. Sci.* **2005**, *33*, 493–530.

(6) Mao, H.-k.; Hemley, R. J. Ultrahigh-pressure transitions in solid hydrogen. *Rev. Mod. Phys.* **1994**, *66*, 671.

(7) Howie, R. T.; Guillaume, C. L.; Scheler, T.; Goncharov, A. F.; Gregoryanz, E. Mixed molecular and atomic phase of dense hydrogen. *Phys. Rev. Lett.* **2012**, *108*, 125501.

(8) Dalladay-Simpson, P.; Howie, R. T.; Gregoryanz, E. Evidence for a new phase of dense hydrogen above 325 gigapascals. *Nature* **2016**, *529*, 63–67.

(9) Eremets, M. I.; Drozdov, A. P.; Kong, P.; Wang, H. Semimetallic molecular hydrogen at pressure above 350 GPa. *Nat. Phys.* **2019**, *15*, 1246–1249.

(10) Loubeyre, P.; Occelli, F.; Dumas, P. Synchrotron infrared spectroscopic evidence of the probable transition to metal hydrogen. *Nature* **2020**, *577*, 631–635.

(11) Ji, C. et al. Ultrahigh-pressure crystallographic passage towards metallic hydrogen. *Nature* **2025**, 1–6.

(12) Pickard, C. J.; Needs, R. J. Structure of phase III of solid hydrogen. *Nat. Phys.* **2007**, *3*, 473–476.

(13) Liu, H.; Wang, H.; Ma, Y. Quasi-molecular and atomic phases of dense solid hydrogen. *J. Phys. Chem. C* **2012**, *116*, 9221–9226.

(14) McMinis, J.; Clay III, R. C.; Lee, D.; Morales, M. A. Molecular to atomic phase transition in hydrogen under high pressure. *Phys. Rev. Lett.* **2015**, *114*, 105305.

(15) Monserrat, B.; Drummond, N. D.; Dalladay-Simpson, P.; Howie, R. T.; López Ríos, P.; Gregoryanz, E.; Pickard, C. J.; Needs, R. J. Structure and metallicity of phase V of hydrogen. *Phys. Rev. Lett.* **2018**, *120*, 255701.

(16) Monacelli, L.; Casula, M.; Nakano, K.; Sorella, S.; Mauri, F. Quantum phase diagram of high-pressure hydrogen. *Nat. Phys.* **2023**, *19*, 845–850.

(17) Nellis, W. J.; Weir, S. T.; Mitchell, A. C. Metallization and Electrical Conductivity of Hydrogen in Jupiter. *Science* **1996**, *273*, 936–938.

(18) Weir, S.; Mitchell, A.; Nellis, W. J. Metallization of fluid molecular hydrogen at 140 GPa (1.4 Mbar). *Phys. Rev. Lett.* **1996**, *76*, 1860.

(19) Loubeyre, P. et al. Coupling static and dynamic compressions: first measurements in dense hydrogen. *High Press. Res.* **2004**, *24*, 25–31.

(20) Knudson, M. D.; Desjarlais, M. P.; Becker, A.; Lemke, R. W.; Cochrane, K.; Savage, M. E.; Bliss, D. E.; Mattsson, T.; Redmer, R. Direct observation of an abrupt insulator-to-metal transition in dense liquid deuterium. *Science* **2015**, *348*, 1455–1460.

(21) Zaghoo, M.; Silvera, I. F. Conductivity and dissociation in liquid metallic hydrogen and implications for planetary interiors. *Proc. Natl. Acad. Sci.* **2017**, *114*, 11873–11877.

(22) Celliers, P. M.; Millot, M.; Brygoo, S.; McWilliams, R. S.; Fratanduono, D. E.; Rygg, J. R.; Goncharov, A. F.; Loubeyre, P.; Eggert, J. H.; Peterson, J. L.; Meezan, N. B.; Le Pape, S.; Collins, G. W.; Hemley, R. J. Insulator-metal transition in dense fluid deuterium. *Science* **2018**, *361*, 677–682.

(23) Scandolo, S. Liquid–liquid phase transition in compressed hydrogen from first-principles simulations. *Proc. Natl. Acad. Sci.* **2003**, *100*, 3051–3053.

(24) Delaney, K. T.; Pierleoni, C.; Ceperley, D. Quantum Monte Carlo Simulation of the High-Pressure Molecular-Atomic Crossover in Fluid Hydrogen. *Phys. Rev. Lett.* **2006**, *97*, 235702.

(25) Morales, M. A.; Pierleoni, C.; Schwegler, E.; Ceperley, D. M. Evidence for a first-order

liquid-liquid transition in high-pressure hydrogen from ab initio simulations. *Proc. Natl. Acad. Sci.* **2010**, *107*, 12799–12803.

(26) Lorenzen, W.; Holst, B.; Redmer, R. First-order liquid-liquid phase transition in dense hydrogen. *Phys. Rev. B* **2010**, *82*, 195107.

(27) Mazzola, G.; Helled, R.; Sorella, S. Phase diagram of hydrogen and a hydrogen-helium mixture at planetary conditions by quantum Monte Carlo simulations. *Phys. Rev. Lett.* **2018**, *120*, 025701.

(28) Cheng, B.; Mazzola, G.; Pickard, C. J.; Ceriotti, M. Evidence for supercritical behaviour of high-pressure liquid hydrogen. *Nature* **2020**, *585*, 217–220.

(29) Tirelli, A.; Tenti, G.; Nakano, K.; Sorella, S. High-pressure hydrogen by machine learning and quantum Monte Carlo. *Phys. Rev. B* **2022**, *106*, L041105.

(30) Dong, X.; Xie, H.; Chen, Y.; Liang, W.; Zhang, L.; Wang, L.; Wang, H. Deep variational free energy prediction of dense hydrogen solid at 1200 K. *Phys. Rev. B* **2025**, *111*, 214118.

(31) Bischoff, T.; Jäckl, B.; Rupp, M. Hydrogen under Pressure as a Benchmark for Machine-Learning Interatomic Potentials. *arXiv:2409.13390* **2024**,

(32) Istas, M.; Jensen, S.; Yang, Y.; Holzmann, M.; Pierleoni, C.; Ceperley, D. M. Liquid-liquid phase transition of hydrogen and its critical point: Analysis from ab initio simulation and a machine-learned potential. *Phys. Rev. E* **2025**, *111*, 045307.

(33) Tenti, G.; Jäckl, B.; Nakano, K.; Rupp, M.; Casula, M. Hydrogen liquid-liquid transition from first principles and machine learning. *Phys. Rev. B* **2025**, *112*, 104208.

(34) Becke, A. D.; Edgecombe, K. E. A simple measure of electron localization in atomic and molecular systems. *J. Chem. Phys.* **1990**, *92*, 5397–5403.

(35) Geng, H. Y.; Wu, Q.; Marqués, M.; Ackland, G. J. Thermodynamic anomalies and three distinct liquid-liquid transitions in warm dense liquid hydrogen. *Phys. Rev. B* **2019**, *100*, 134109.

(36) de Villa, K.; Wang, X.; Zurek, E.; Militzer, B. Superionicity in ammonium polyhydrides at extreme pressures. *J. Chem. Phys.* **2025**, *163*.

(37) Belli, F.; Novoa, T.; Contreras-García, J.; Errea, I. Strong correlation between electronic bonding network and critical temperature in hydrogen-based superconductors. *Nat. Commun.* **2021**, *12*, 5381.

(38) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*, 184115.

(39) Behler, J.; Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.

(40) Riffet, V.; Labet, V.; Contreras-García, J. A topological study of chemical bonds under pressure: solid hydrogen as a model case. *Phys. Chem. Chem. Phys.* **2017**, *19*, 26381–26395.

(41) Novoa, T.; Di Mauro, M. E.; Inostroza, D.; El Haloui, K.; Sisourat, N.; Maday, Y.; Contreras-García, J. TcESTIME: predicting high-temperature hydrogen-based superconductors. *Chemical Science* **2025**, *16*, 57–68.

(42) Hornik, K.; Stinchcombe, M.; White, H. Multilayer feedforward networks are universal approximators. *Neural Netw.* **1989**, *2*, 359–366.

(43) Paszke, A. et al. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*.

(44) Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* **2017**,

(45) Huber, P. J. *Breakthroughs in statistics: Methodology and distribution*; Springer, 1992; pp 492–518.

(46) LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **2002**, *86*, 2278–2324.

(47) Fang, Y.-W.; Errea, I. Superconductivity in barium hydrides via incorporation of light elements. *Phys. Rev. B* **2025**, *112*, 125204.

(48) Belli, F.; Torres, S.; Contreras-García, J.; Zurek, E. Refining Tc Prediction in Hydrides via Symbolic-Regression-Enhanced Electron-Localization-Function-Based Descriptors. *Ann. Phys.* **2025**, e00280.

(49) Avery, P.; Zurek, E. RandSpg: An open-source program for generating atomistic crystal structures with specific spacegroups. *Comput. Phys. Commun.* **2017**, *213*, 208–216.

(50) Falls, Z.; Avery, P.; Wang, X.; Hilleke, K. P.; Zurek, E. The XtalOpt evolutionary algorithm for crystal structure prediction. *J. Phys. Chem. C* **2020**, *125*, 1601–1620.

(51) Zhao, W.; Huang, X.; Zhang, Z.; Chen, S.; Du, M.; Duan, D.; Cui, T. Superconducting ternary hydrides: progress and challenges. *Natl. Sci. Rev.* **2024**, *11*, nwad307.

(52) Tsirelson, V.; Stash, A. Determination of the electron localization function from electron density. *Chem. Phys. Lett.* **2002**, *351*, 142–148.

(53) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **1996**, *77*, 3865.

(54) Blochl, P. E. Projector augmented-wave method. *Phys. Rev. B* **1994**, *50*, 17953.

(55) Kresse, G.; Hafner, J. *Ab initio* molecular dynamics for liquid metals. *Phys. Rev. B.* **1993**, *47*, 558.

(56) Baldereschi, A. Mean-value point in the Brillouin zone. *Phys. Rev. B* **1973**, *7*, 5212.

(57) Nosé, S. A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.* **1984**, *81*, 511–519.

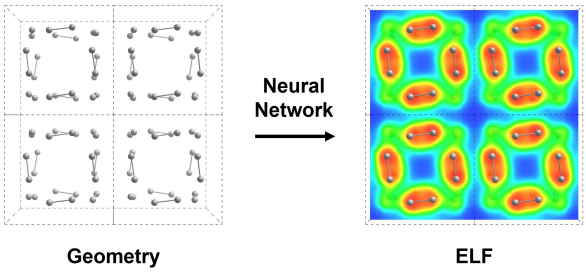
(58) Hoover, W. G. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A* **1985**, *31*, 1695.

(59) Otero-de-la Roza, A.; Johnson, E. R.; Luaña, V. Critic2: A program for real-space analysis of quantum chemical interactions in solids. *Comput. Phys. Commun.* **2014**, *185*, 1007–1018.

Supporting Information

Supporting Information is available and includes: evaluation of the convolutional neural network (CNN) model; hyperparameter convergence tests; extended error analysis; and a supplementary table containing structure-wise statistics for randomly generated structures.

TOC Graphic



We present a machine-learning framework that predicts the electron localization function (ELF) of dense hydrogen directly from atomic geometry, bypassing explicit electronic-structure calculations. Trained on ab initio data for fluid hydrogen across multiple pressures, the model achieves high accuracy and reveals pressure-dependent nonlocal contributions, while transferring robustly to crystalline hydrogen and preserving key ELF topological features.

Supporting Information for Geometry-Based Neural-Network Prediction of Electron Localization Function Topology in Dense Hydrogen

Xiaoyu Wang,[*,†] Miriam Marqués,[‡] Sergio Gómez,[¶,§] Francesc Serratosa,[¶] Eva Zurek,[‖] and Julia Contreras-García[*,†]

†*Sorbonne Université, CNRS, Laboratoire de Chimie Théorique, LCT, 75005 Paris, France*

‡*CSEC, School of Physics and Astronomy, University of Edinburgh, Edinburgh, EH9 3JZ, United Kingdom*

¶*Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, 43007 Tarragona, Spain*

§*ComSCIAM, Universitat Rovira i Virgili, 43007 Tarragona, Spain*

‖*Department of Chemistry, State University of New York at Buffalo, Buffalo, NY 14260-3000, USA*

E-mail: xiaoyu.wang@sorbonne-universite.fr;

julia.contreras_garcia@sorbonne-universite.fr

Contents

S1 Test on Convolutional Neural Network

We have additionally tested a single-channel three-dimensional convolutional neural network[1] (3D CNN) as an alternative model architecture to assess whether non-local correlations can be captured within a grid-based framework. For each sampling point $\mathbf{r}_0$, a local density patch is constructed as

$$\rho(\mathbf{x}) = \sum_{i \in r_{\mathrm{cut}}} \exp\left(-\frac{|\mathbf{x} - (\mathbf{R}_i - \mathbf{r}_0)|^2}{2\sigma^2} \right), \tag{1}$$

where the sum runs over hydrogen atoms within a cutoff radius r_{cut}. The density is discretized on a cubic grid spanning $[-r_{\mathrm{cut}}, r_{\mathrm{cut}}]^3$.

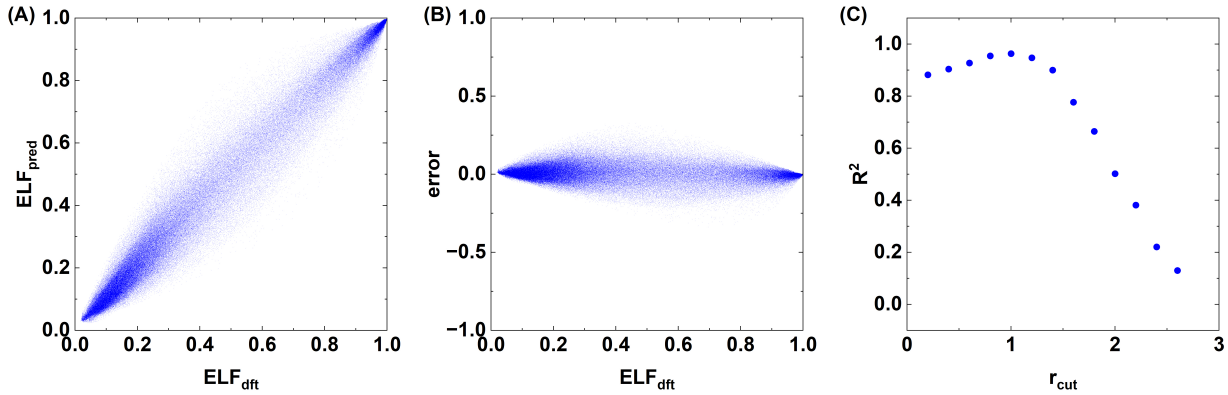


Figure S1: (A) Correlation between predicted and reference ELF values; (B) Prediction error ($\mathrm{ELF}_{\mathrm{pred}} - \mathrm{ELF}_{\mathrm{true}}$) as a function of the reference ELF; (C) Dependence of the coefficient of determination (R^2) on the cutoff radius r_{cut}.

The resulting input tensor is

$$\mathbf{X} \in \mathbb{R}^{N \times P^3}, \tag{2}$$

where N is the number of sampled grid points and P is the number of discretization points per spatial dimension. In this work, we set $P = 12$ and used a CNN consisting of three convolutional blocks followed by a sigmoid output layer. Using the same training protocol as for the MLP model[2] (50,000 samples per MD snapshot), we systematically varied the cutoff radius r_{cut} from 0.25 to 2.5 Å. The best performance was obtained at $r_{\mathrm{cut}} = 1.0$ Å,

yielding a maximum coefficient of determination $R^2 = 0.966$ for the 150,000 evaluation set.

Despite its increased representational flexibility, the CNN model exhibits several practical limitations. First, the input dimensionality becomes prohibitively large (on the order of 10^8 elements for the present setup), which severely restricts model scalability. Second, the grid-based representation is inherently tied to a three-dimensional voxelization and is therefore not directly transferable to lower-dimensional analyses (e.g., line scans or planar slices), which are commonly used in ELF studies.

Importantly, the spatial structure of the prediction residuals remains qualitatively similar to that obtained with the MLP model, albeit with a broader distribution. In particular, the residual retains a pronounced long-wavelength character, indicating that the dominant source of error is not specific to the choice of model architecture. We further observe that increasing r_{cut} does not improve performance beyond the optimal value. While a larger cutoff formally incorporates more long-range information, it also dilutes the resolution of short-range features due to the fixed grid size, leading to a decrease in predictive accuracy. This trade-off results in a clear optimum at $r_{\mathrm{cut}} = 1.0$ Å. Overall, these results suggest that the observed long-wavelength residual is not primarily a consequence of the specific local regression model, but rather reflects intrinsic limitations in representing long-range correlations within a strictly local descriptor framework.

S2 Convergence Tests of Hyperparameters

We performed systematic convergence tests with respect to the key hyperparameters of the model. Unless otherwise specified, the baseline configuration consists of a network width of 128, two hidden layers, a cutoff radius $r_{\mathrm{cut}} = 3.0$, $n_{\mathrm{radial}} = 10$, $l_{\mathrm{max}} = 2$, and a Huber loss parameter $\delta = 0.03$, trained for 80 epochs.

To ensure a consistent comparison across different hyperparameter settings, the random seeds used to generate the training set ($50{,}000 \times 3$ samples) and the evaluation set ($150{,}000 \times 3$ samples) were kept fixed throughout all tests.

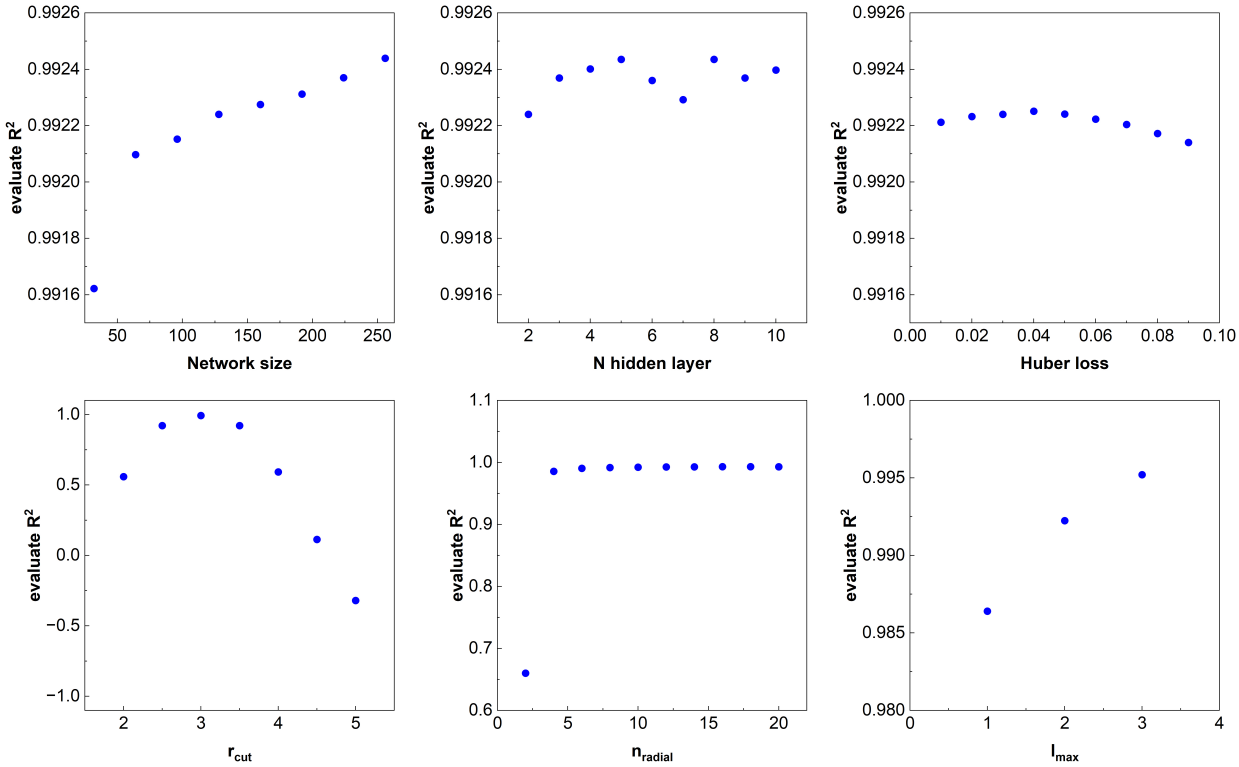


Figure S2: Convergence of model performance with respect to key hyperparameters. The coefficient of determination (R^2) on the validation set is shown as a function of (i) network width, (ii) number of hidden layers, (iii) Huber loss parameter, (iv) cutoff radius r_{cut}, (v) number of radial basis functions n_{radial}, and (vi) maximum angular momentum l_{max}.

The impact of numerical precision was assessed by repeating the evaluation using float32 storage, yielding differences in validation R^2 below 10^{-5}, indicating that float16 precision does not introduce measurable artifacts.

For a representative 500-atom hydrogen structure on a 192^3 grid, the ML pipeline (descriptor construction and inference) requires approximately 5–8 seconds on a single NVIDIA V100 GPU.

S3 Extra Residue Analysis

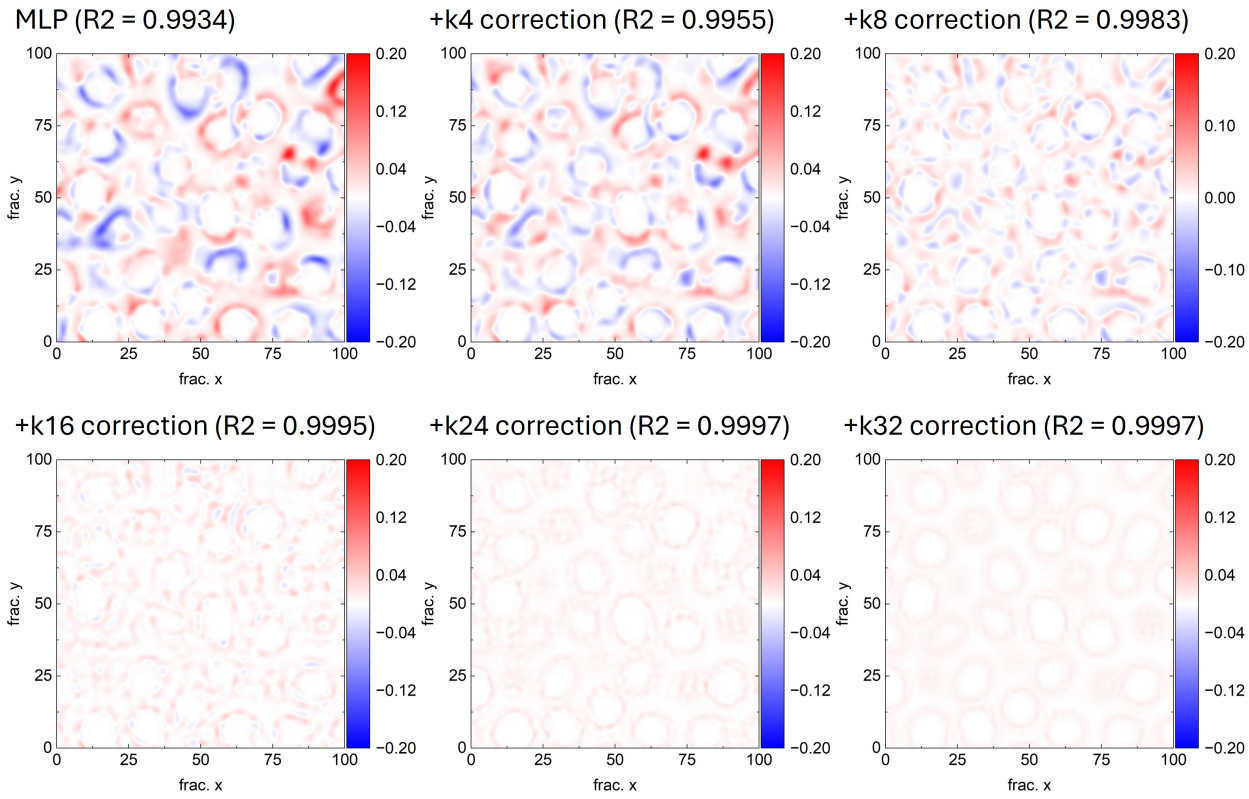


Figure S3: Real-space residuals between predicted and DFT ELF values for a representative two-dimensional slice of the 76.0 GPa structure. The color scale ranges from -0.2 (blue), through 0 (white), to $+0.2$ (red). Results are shown for the baseline model and for predictions after applying different k-correction schemes, in which the residual in logit space is projected onto a truncated low-k Fourier basis to restore long-wavelength contributions beyond the local descriptor.

S4 Advanced Error Analysis

To quantify the uncertainty of the reported metrics, we repeated the entire workflow—including training set generation, model training, and validation set sampling—50 times using different random seeds, resulting in an ensemble of 50 independently trained models. For ELF training and validation, the resulting standard deviations of MAE, RMSE, and R^2 are on the order of 10^{-5}, indicating negligible stochastic uncertainty and highly stable model performance.

The same ensemble was then applied to ELF prediction and networking value (NW) prediction on randomly generated structures using RandSpg.[3] In this case, the global standard deviation across all structures is significantly larger; however, this spread is dominated by genuine variation in performance among different structures rather than training stochasticity. Accordingly, we analyze the results on a per-structure basis, where the run-to-run variability remains small and the variation of the structure-wise mean values reflects differences in structural complexity. All individual metrics for the 50 runs, including both ELF and NW predictions, are provided in an accompanying Excel file for completeness.

Table S1: Mean and standard deviation of ELF prediction metrics and networking value (NW) predictions for training data and randomly generated (RandSpg) cubic and hexagonal hydrogen structures. Statistics are computed from an ensemble of 50 independently trained models and reported globally (across all samples).

	mean MAE	stdev	mean RMSE	stdev	mean R^2	stdev
Training						
MLP	0.0190	6.83×10^{-5}	0.0271	9.06×10^{-5}	0.992	5.39×10^{-5}
RandSpg (ELF)						
Cubic	0.0385	0.0159	0.0520	0.0196	0.956	0.0421
Hexagonal	0.0470	0.0148	0.0635	0.0200	0.947	0.0407
RandSpg (NW)						
Cubic	0.0331	0.0243	0.0372	0.0246	0.940	-
Hexagonal	0.0459	0.0305	0.0506	0.0304	0.866	-

To further validate the model transferability, we performed an additional test on a completely independent set of snapshots taken from AIMD trajectories. At each pressure point

(76.0, 115.1, and 138.5 GPa), two snapshots were selected at 1000 and 2000 MD steps, corresponding to 0.5 and 1.0 ps after equilibration. Following the same procedure as in the main manuscript, 150,000 random voxels were sampled from each snapshot. The model was not retrained; instead, the same model used for the production calculations in the manuscript was directly applied. The results are summarized in Table S2

Table S2: Prediction performance (MAE, RMSE, and R^2) for independent AIMD snapshots (1000 and 2000 steps; 0.5 and 1.0 ps after equilibration) at each pressure point.

Pres.	Snapshot	MAE	RMSE	R^2
76.0	1000	0.0194	0.0265	0.991
76.0	2000	0.0197	0.0268	0.991
115.1	1000	0.0216	0.0305	0.991
115.1	1000	0.0211	0.0298	0.991
138.5	1000	0.0185	0.0279	0.993
138.5	2000	0.0185	0.0279	0.993

References

(1) LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **2002**, *86*, 2278–2324.

(2) Hornik, K.; Stinchcombe, M.; White, H. Multilayer feedforward networks are universal approximators. *Neural Netw.* **1989**, *2*, 359–366.

(3) Avery, P.; Zurek, E. RandSpg: An open-source program for generating atomistic crystal structures with specific spacegroups. *Comput. Phys. Commun.* **2017**, *213*, 208–216.