Supplementary Materials

Prediction of electron localization functions from superposed atomic densities for accelerated superhydride discovery

Austin Ellis[*], Samantha Ellis, Maosheng Miao[*]

Department of Chemistry and Biochemistry, California State University, Northridge, CA 91330, USA.

[*]**Correspondence to:** Austin Ellis, Prof. Maosheng Miao, Department of Chemistry and Biochemistry, California State University, Northridge, CA 91330, USA. E-mail: ellisal@ornl.gov; mmiao@csun.edu

This Supplementary Material contains twelve tables, one methodological note, and one training-dynamics figure supporting the main manuscript.

Supplementary Note 1	Pressure-targeted lattice expansion
Supplementary Tables 1–12	Dataset, evaluation-set, timing, and pressure-targeting summaries
Supplementary Figure 1	Production-checkpoint training dynamics

Supplementary Note 1: Pressure-Targeted Lattice Expansion

The high-pressure training set combines an initial fixed-lattice grid with a second pressure-targeted multiplier grid. The purpose of the multiplier stage was to increase coverage of physically relevant compressed volumes without introducing new prototype shapes, new internal-coordinate degrees of freedom, or additional structural relaxation steps. In this stage, the source lattice-vector directions and fractional atomic coordinates were preserved, while the lattice-vector lengths were rescaled to target values inferred from pressure–scale trends.

Fixed-lattice seed calculations were organized by structure class, prototype, chemical system, and scale token. Only calculations classified as self-consistent-field (SCF) converged were used to construct pressure–scale relations. Final pressures were parsed from the Vienna Ab initio Simulation Package (VASP) stress output and converted to GPa; these are achieved DFT pressures of the fixed or rescaled cells, not externally imposed PSTRESS values. Within each group, the median final pressure was assigned to each scale value in order to reduce the influence of noisy or anomalous individual calculations.

For each grouped system g, let s denote the scalar lattice scale and let $P_g(s)$ denote the median achieved pressure at that scale. Because pressure is expected to decrease as the lattice scale increases, the median pressure–scale points were fit with a monotonic non-increasing curve,

$$\widetilde{P}_g(s_1) \geq \widetilde{P}_g(s_2) \qquad \text{for} \qquad s_1 < s_2, \tag{S1}$$

using isotonic regression. This step denoises the fixed-lattice pressure data while enforcing the physically expected ordering needed to invert the curve.

A reference scale $a_{0,g}$ was then estimated for each system where the smoothed curve crossed zero pressure:

$$\widetilde{P}_g(a_{0,g}) = 0. \tag{S2}$$

The usable target-pressure interval for a system was defined as the overlap between the pressure range covered by $\widetilde{P}_g$ and the project target interval $[0,500]$ GPa. Seven target pressures were then placed uniformly across this usable interval. For systems spanning the full target range, these targets are

$$P_t \in \{0,\ 83.333,\ 166.667,\ 250,\ 333.333,\ 416.667,\ 500\}\ \text{GPa}. \tag{S3}$$

The monotonic fitted curve was inverted to obtain target scales a_t satisfying

$$\widetilde{P}_g(a_t)=P_t, \tag{S4}$$

and the corresponding multiplier was computed as

$$m_t=\frac{a_t}{a_{0,g}}. \tag{S5}$$

For anisotropic cells, per-axis lattice ratios were estimated from the fixed-lattice seed structures by dividing each lattice-vector length by the scale token and taking medians over the available scale points. Target structures were generated by preserving the lattice-vector directions and internal fractional coordinates of the source structure, while rescaling the lattice-vector lengths according to the pressure-targeted scale. Thus, the multiplier stage increases pressure coverage while keeping the topology of each compact prototype fixed. This makes the additional data well matched to the learning objective: predicting how ELF basins move, sharpen, weaken, or disappear as prototype lattices are compressed.

Supplementary Table 1. Element inventory of the final high-pressure PBE dataset used for electron localization function (ELF) learning.

Quantity	**Value**
Number of elements	89
Included elements	H, He, Li, Be, B, C, N, O, F, Ne, Na, Mg, Al, Si, P, S, Cl, Ar, K, Ca, Sc, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Ga, Ge, As, Se, Br, Kr, Rb, Sr, Y, Zr, Nb, Mo, Tc, Ru, Rh, Pd, Ag, Cd, In, Sn, Sb, Te, I, Xe, Cs, Ba, La, Ce, Pr, Nd, Pm, Sm, Eu, Gd, Tb, Dy, Ho, Er, Tm, Yb, Lu, Hf, Ta, W, Re, Os, Ir, Pt, Au, Hg, Tl, Pb, Bi, Ac, Th, Pa, U, Np, Pu

Supplementary Table 2. Pressure-targeting and convergence accounting for the multiplier expansion stage. SCF denotes self-consistent field.

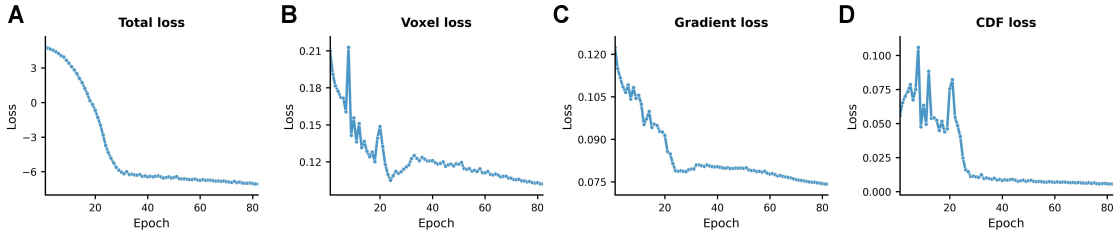
Quantity	Count or value
Systems analyzed for pressure targeting	27,944
Systems with usable pressure-scale relations	27,813
Systems excluded	131
Excluded because of insufficient pressure–scale points	16
Excluded because of no overlap with the target-pressure window	115
Candidate calculations evaluated	472,804
SCF converged calculations	425,285
Multiplier range	0.3500-1.0714
Multiplier median	0.7481
Final fixed-lattice structures	161,876
Final pressure-targeted structures	164,133
Final selected structures	326,009

Supplementary Table 3. Median discrepancies for the 96-structure exchange-correlation-functional cross-check. The subset contained 16 structures at each of 50, 100, 160, 200, 250, and 300 GPa. PBE and r²SCAN ELF fields were calculated on identical fixed geometries and grids.

Comparison	Median voxel MAE	Median maximum-value error	Median maximum-location distance (Å)
ELFNet vs PBE	0.09871	0.02954	1.1094
ELFNet vs r²SCAN	0.10493	0.03714	1.1832
PBE vs r²SCAN	0.01322	0.00494	0.0000

Supplementary Table 4. Summary of the final high-pressure Perdew–Burke–Ernzerhof (PBE) dataset used for electron localization function (ELF) learning.

Quantity	Value
Total structures	326,009
Total atoms	709,017
Unary structures	7,423
Binary structures	318,586
Elements represented	89
Achieved-pressure window	−49.983 to 499.997 GPa
Median achieved pressure	83.26 GPa
Mean achieved pressure	149.12 GPa
Maximum achieved pressure	499.997 GPa
Cohesive energy per atom, median	−1.341 eV atom^{-1}
Force magnitude, median	0.00 eV Å^{-1}
Force magnitude, mean	12.84 eV Å^{-1}



Supplementary Figure 1. Training dynamics for the production ELFNet checkpoint. The panels show total loss, voxel loss, gradient loss, and cumulative-distribution-function (CDF) loss as functions of epoch.

Supplementary Table 5. Inference-time symmetry-averaging results for the 117 Fm-3m structures in the evaluation set. Negative changes in DFT-referenced errors indicate improvement after symmetry averaging.

Quantity	Median
Raw-versus-symmetry-averaged field MAE	0.01672
Change in predicted maximum ELF value	0.00740
Change in single-voxel argmax location	0.640 Å
Change in DFT-referenced voxel MAE	−0.00245
Change in DFT-referenced maximum-location error	−0.185 Å

Note. The 0.640 Å value reflects near-degenerate peak-voxel switching, not physical displacement of an ELF feature.

Supplementary Table 6. Matched-budget architecture ablations and plain three-dimensional U-Net baseline over the 50,000-structure evaluation set. All models were trained for 60 epochs. The convolution multiply-accumulate (MAC) ratio is normalized to production ELFNet. Values are from the common-source evaluation.

(a) Model configuration and compute

Model or variant	Active loss	Parameters	MAC ratio
Full ELFNet	Full composite	4,232,989	1.000
ELFNet without CBAM	Full composite	4,232,261	1.000
ELFNet with zero padding	Full composite	4,232,989	1.000
FlatResNet3D without peak loss	Voxel + gradient + CDF	4,232,988	1.000
Plain 3D U-Net	Voxel + gradient + CDF	4,442,092	0.939

(b) Evaluation performance

Model or variant	Joint pass (%)	Median max-value error	Median max-location distance (Å)	Median voxel MAE
Full ELFNet	92.424	0.02680	0.9156	0.08697
ELFNet without CBAM	75.904	0.03357	1.3003	0.10040
ELFNet with zero padding	70.006	0.05852	1.7791	0.13668
FlatResNet3D without peak loss	76.432	0.03050	1.1855	0.08847
Plain 3D U-Net	71.840	0.06527	1.3561	0.11350

Supplementary Table 7. Composite-loss ablation over the 50,000-structure evaluation set. All variants used the same architecture and 60-epoch training schedule. Joint pass requires maximum-value error ≤ 0.20 and maximum-location distance ≤ 3 Å. Values are from the common-source evaluation.

Active loss terms	Joint pass (%)	Median max-value error	Median max-location distance (Å)	Median voxel MAE
Voxel only	77.564	0.03250	1.2016	0.08428
Voxel + gradient	77.564	0.03421	1.2204	0.08302
Voxel + gradient + CDF	76.432	0.03050	1.1855	0.08847
Voxel + gradient + CDF + peak	92.424	0.02680	0.9156	0.08697

Supplementary Table 8. Composition and parent-pressure distribution of the selected 50,000-structure hydride-derived metal-template analysis set.

(a) Set composition

Set composition	Count
Total structures	50,000
Binary branch	48,438
Mono-element branch	1,562
Parent prototype labels	55
Parent hydride formulas	50
Distinct space groups	28
Unique element-assignment labels	1,890

(b) Parent-pressure distribution

Parent pressure	Structures	Fraction
50 GPa	2,172	4.34%
100 GPa	1,228	2.46%
160 GPa	811	1.62%
200 GPa	42,102	84.20%
250 GPa	2,078	4.16%
300 GPa	1,609	3.22%

Supplementary Table 9. Parent hydride formula-to-prototype mapping for the selected 50,000-structure hydride-derived metal-template analysis set. The set contains 50 formulas and 55 prototype labels.

Parent hydride formula	Associated prototype label(s)
AcH10	AcH10_Cm_200
AcH16	AcH16_P6mmm_200
AcH4	AcH4_Cmcm_200
AcH5	AcH5_C2m_50GPa_200
AcH7	AcH7_Cmc21_200
AcH8	AcH8_C2m_200
BaH10	BaH10_Cmmm_200
BaH12	BaH12_I4mmm_250GPa_200; BaH12_P21_200; BaH12_P63mmc_200
BaH6	BaH6_Imm2_200
BaH8	BaH8_Imma_200
CaH12	CaH12_C2c_200GPa_200; CaH12_R-3_100GPa_200
CaH6	CaH6_Im-3m_200
CaH9	CaH9_C2m_300GPa_200
CeH9	CeH9_P63mmc_200
HfH14	HfH14_C2m_200
HfH4	HfH4_Fddd_200
HfH6	HfH6_Cmc21_200
LaH11	LaH11_P4nmm_200
LaH16	LaH16_P6mmm_200
LaH4	LaH4_I4mmm_200
LaH6	LaH6_R-3c_200
LaH7	LaH7_Cmc21_50GPa_200
LaH9	LaH9_Cc_50GPa_200
NdH9	NdH9_F-43m_200
PaH4	PaH4_Fmmm_200
PrH6	PrH6_C2m_200
PrH8	PrH8_P63mc_200

Parent hydride formula	Associated prototype label(s)
ScH10	ScH10_Cmcm_250GPa_200
ScH12	ScH12_C2c_300GPa_200; ScH12_Cmcm_200
ScH6	ScH6_Cmcm_200
ScH7	ScH7_Cmcm_200
ScH8	ScH8_Immm_200
ScH9	ScH9_I41md_300GPa_200
SrH10	SrH10_P21c_200
SrH12	SrH12_C2m_200
SrH8	SrH8_P21c_200
ThH10	ThH10_Fm-3m_200
ThH7	ThH7_P21c_200
Ti2H13	Ti2H13_Immm_200
TiH14	TiH14_C2m_200
TiH16	TiH16_55_200
TiH22	TiH22_C2m_200
UH7	UH7_P63mmc_200
UH8	UH8_Fm-3m_200
YH12	YH12_C2m_200
YH13	YH13_R-3m_200GPa_200
YH9	YH9_P63m_100GPa_200
ZrH10	ZrH10_P63mmc_200
ZrH6	ZrH6_Cmc21_160GPa_200; ZrH6_P21c_200
ZrH8	ZrH8_I4mmm_200

Supplementary Table 10. Maximum-value and maximum-location errors and screening-threshold regions for the selected 50,000-structure hydride-derived metal-template analysis set.

(a) Overall performance

Quantity	Value
Structures analyzed	50,000
Median maximum-value absolute error	0.0268
Maximum-value absolute error 95th percentile	0.1577
Median maximum-location distance	0.895 Å
Maximum-location distance 95th percentile	2.845 Å

(b) Screening-threshold regions

Maximum-value error	Maximum-location distance	Structures	Fraction
≤ 0.20	≤ 3 Å	46,786	93.57%
≤ 0.20	>3 Å	1,699	3.40%
> 0.20	≤ 3 Å	1,515	3.03%
> 0.20	>3 Å	0	0.00%

Supplementary Table 11. Production-model performance and failure rate stratified by parent pressure over the 50,000-structure evaluation set. Failure denotes violation of either the maximum-value-error threshold of 0.20 or maximum-location-distance threshold of 3 Å.

Pressure (GPa)	n	Median maximum-value error	Median maximum-location distance (Å)	Failure rate (%)
50	2,172	0.03147	1.0140	10.13
100	1,228	0.03001	0.9841	2.36
160	811	0.02107	1.2957	13.56
200	42,102	0.02647	0.8931	6.14
250	2,078	0.02576	0.7312	3.85
300	1,609	0.03602	0.8245	11.93

Supplementary Table 12. Timing definitions for fixed-geometry DFT ELF generation and ELFNet-predicted ELFCAR generation over 50,000 structures.

Quantity	Value
Sum of VASP "Total CPU time used" fields	922.6 h
Mean VASP CPU-timer value per structure	66.43 s
Median VASP CPU-timer value per structure	30.26 s
VASP CPU-timer 95th percentile	235.55 s
Sum of VASP "Elapsed time" fields	935.3 h
Mean VASP elapsed time per structure	67.34 s
Sum of ELFNet pipeline elapsed times	2.67 h
Mean ELFNet pipeline elapsed time per structure	0.192 s
Median ELFNet pipeline elapsed time per structure	0.153 s
ELFNet pipeline elapsed-time 95th percentile	0.441 s
Aggregate model forward-pass time	18.5 min
Mean model forward-pass time per structure	0.0222 s
Ratio of summed elapsed times (VASP/ELFNet)	350.3×

Note. VASP values sum OUTCAR timers, not scheduler core-hours; ELFNet sums per-structure elapsed time across four GPUs. The 350.3× ratio is hardware- and concurrency-specific, not a campaign wall-clock or hardware-normalized speedup.