ELECTRON-DENSITY-INFORMED MOLECULAR PROPERTY PREDICTION USING SURROGATE FUNCTIONALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine learning methods are increasingly used to predict chemical properties directly from the molecular structure to bypass the often prohibitive cost of quantum mechanical calculations. Recent work has shown that including the ground-state electron density as input to the machine learning model can improve accuracy in this setting. This work employs the recent surrogate functional framework originally developed for orbital-free density functional theory and extends it to predicting molecular properties. We show that passing the ground-state density as input to the model, while simultaneously enforcing it to be the minimum of the surrogate functional, surpasses state-of-the-art performance for six out of seven extensive properties of the QM9 dataset. Additionally, the approach shows improved generalization to larger molecules taken from the QM40 dataset. On all extensive properties of both datasets, our approach outperforms models that are informed by the ground-state density but do not enforce the surrogate functional constraints.

1 INTRODUCTION

Predicting molecular properties from first principles is a central goal of computational chemistry. However, quantum-mechanical methods such as Kohn–Sham density functional theory (DFT) (Kohn & Sham, 1965) remain too computationally expensive for many applications. Machine learning circumvents this cost by learning the mapping from molecular structure to properties directly from reference calculations (von Lilienfeld et al., 2020; Musil et al., 2021; Keith et al., 2021). In particular, graph neural networks (Schütt et al., 2018; Anstine et al., 2025), including equivariant architectures (Schütt et al., 2021; Batzner et al., 2022; Batatia et al., 2022; Fu et al., 2025; Aykent & Xia, 2025), have achieved remarkable accuracy on this task.

A complementary approach is to incorporate information about the electronic structure of the molecule. By the first Hohenberg–Kohn theorem (Hohenberg & Kohn, 1964), the ground-state electron density uniquely determines the external potential, up to an additive constant, and thus, in principle, all ground-state properties of the system. The density therefore provides a meaningful physical descriptor for molecular property prediction, and several approaches have used it as input to learned property models (Brockherde et al., 2017; Bogojeski et al., 2020; 2026).

Surrogate functionals (Remme & Hamprecht, 2026; Wagner et al., 2026), originally introduced in the context of machine-learned orbital-free DFT, go beyond treating the density merely as an input representation. Instead, they construct a learned functional whose minimum is attained at the ground-state density and whose value there equals the corresponding ground-state energy. This variational structure yields energy prediction accuracy competitive with state-of-the-art interatomic potentials while substantially improving extrapolation to larger systems (Wagner et al., 2026).

Motivated by these results, we generalize the surrogate functional approach to arbitrary scalar molecular properties. For each target property, we train a dedicated surrogate functional using the same training pipeline, while replacing the ground-state energy with the corresponding property value as the training target. At inference time, the required electron density is supplied by a separately trained network that predicts it directly from molecular structure, avoiding a computationally expensive variational optimization.

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

Our contributions include:

1. We extend the surrogate functional framework from the electronic energy to arbitrary scalar molecular properties.
2. On QM9 (Ramakrishnan et al., 2014), surrogate functionals surpass the previous state of the art on six out of seven extensive properties, with competitive accuracy on the seventh.
3. In a controlled comparison using a common backbone, surrogate functionals consistently outperform both geometry-only models and models that additionally take the electron density as input across all extensive properties. This advantage becomes even more pronounced in the out-of-distribution regime, when models trained exclusively on QM9 are evaluated on the larger molecules in QM40 (Madushanka et al., 2024).
4. We introduce beSEN, a modification of eSEN (Fu et al., 2025) with bilinear activation based on the Gram matrix, which further improves predictive performance, particularly when used within the surrogate functional framework.

2 Background and related work

Molecular property prediction. Most machine-learned models for molecular properties map nuclear charges and coordinates directly to the quantity of interest. Starting from invariant message-passing networks (Schütt et al., 2018; Gasteiger et al., 2022; Liu et al., 2021) and progressing through increasingly expressive equivariant architectures (Schütt et al., 2021; Batzner et al., 2022; Batatia et al., 2022; Thölke & Fabritiis, 2021; Le et al., 2022; Liao & Smidt, 2022; Du et al., 2023; An et al., 2023; Aykent & Xia, 2023; Liao et al., 2024; Fu et al., 2025; Aykent & Xia, 2025), this line of work has reduced energy errors on the QM9 benchmark (Ramakrishnan et al., 2014) to a few meV relative to the reference calculations. In contrast, several models make the electronic structure explicit by using the electron density as an intermediate representation. Brockherde et al. (2017) learn the map from a smoothed external potential to a density expansion, and then fit a second, independent model from the predicted density to the energy. Bogojeski et al. (2020) reuse this two-map construction to learn a coupled-cluster correction on top of a DFT density, and Shao et al. (2023) extend it to the one-electron reduced density matrix. DenSNet (Bogojeski et al., 2026) uses an equivariant network to predict the coefficients of a flexible atom-centered Gaussian basis, while a second equivariant network consumes these coefficients together with the nuclear positions to predict the total energy. DensIP (Wing et al., 2026) follows the same staged recipe but replaces the second network with a four-parameter physics-based intermolecular potential, in the spirit of earlier models built on machine-learned multipoles and valence-density parameters (Bereau et al., 2018).

Machine-learned orbital-free DFT. Orbital-free density functional theory (Mi et al., 2023) obtains the electronic ground state by minimizing an energy functional directly with respect to the density, without ever constructing Kohn–Sham orbitals (Kohn & Sham, 1965). Recent machine-learned functionals have demonstrated energy predictions with chemical accuracy for small drug-like molecules (Zhang et al., 2024; Remme et al., 2025). Surrogate functionals (Remme & Hamprecht, 2026; Wagner et al., 2026) are trained without requiring physical energies at off-equilibrium densities, which allows training from ground-state labels alone, and have been demonstrated to exhibit improved generalization compared to directly predicting the ground-state energy.

3 METHODS

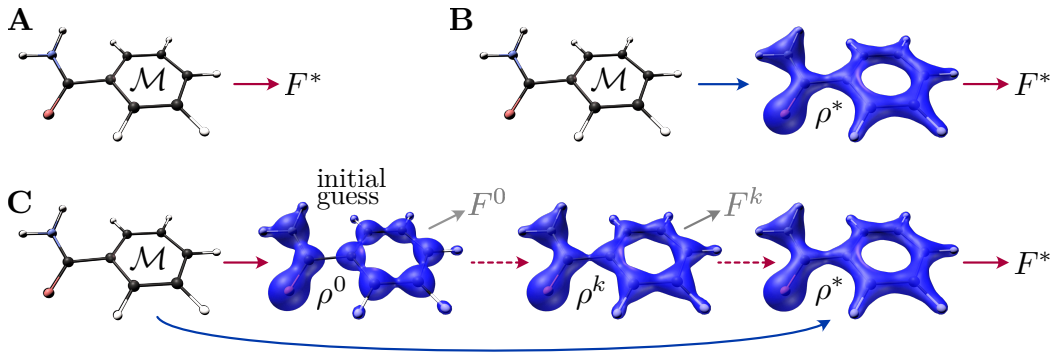

Figure 1: Schematic overview of the three compared frameworks for molecular property prediction. **A** Geometry-only models predict the property directly from the molecular geometry. **B** Density-input models predict the property from the geometry in two steps via the ground-state electron density. **C** Surrogate functional models predict the property from the geometry and the ground-state density, while enforcing that the density is the minimum of a learned surrogate functional. This variational structure makes it possible to reach the ground-state density and property from an initial guess (magenta arrows). At inference time, we use the shortcut of predicting the ground-state density directly with a second network (blue arrow).

3.1 DENSITY REPRESENTATION

Following recent work in machine-learned OF-DFT Zhang et al. (2024); Remme et al. (2025); Remme & Hamprecht (2026), we choose to express the electron density as a linear expansion in atom-centered Gaussian-type basis functions Grisafi et al. (2019); Vergara-Beltran & Rodríguez (2023), to which symmetric Löwdin orthonormalization (Löwdin, 1950) is applied, yielding

$$\rho(\mathbf{r}) = \sum_{\mu} p_\mu \, \omega_\mu(\mathbf{r}) \,, \tag{1}$$

with density coefficients p_μ and orthonormal basis functions $\omega_\mu(\mathbf{r})$. We denote by $\boldsymbol{p}^*$ the coefficients of the ground-state density ρ^*. Each basis function is associated with a definite angular momentum l, so that the coefficients of an atom transform as a direct sum of irreducible representations of $\mathrm{SO}(3)$ and can be used directly as node features of an equivariant message-passing network.

3.2 ORBITAL-FREE SURROGATE FUNCTIONALS

An orbital-free density functional $E[\rho]$ maps a given electron density ρ onto a corresponding energy E. The ground-state values for both the energy and the density can be obtained by minimizing the functional over ρ, which in the expansion of Equation 1 amounts to minimizing $E(\boldsymbol{p})$ over the density coefficients

$$\boldsymbol{p}^* = \arg\min_{\boldsymbol{p}} E(\boldsymbol{p}), \qquad E^* = E(\boldsymbol{p}^*). \tag{2}$$

Machine-learned density functionals (Zhang et al., 2024; Remme et al., 2025) learn an approximation to parts of the physical functional from Kohn–Sham reference data. In contrast, a surrogate functional $E[\boldsymbol{p}]$ (Remme & Hamprecht, 2026; Wagner et al., 2026) is not required to reproduce physical energies at off-equilibrium densities but only needs to converge to the correct minimum during the iterative variational optimization procedure, where Equation 2 is solved via gradient descent. To ensure a stable convergence during inference, surrogate functionals are trained with a train-time variational optimization scheme: For each molecule $\mathcal{M}$, a cached coefficient vector $\boldsymbol{p}_\mathcal{M}$ is maintained and updated each time the molecule appears in a batch via a gradient descent step

$$\boldsymbol{p}_\mathcal{M} \mapsto \boldsymbol{p}_\mathcal{M} - \gamma \, \nabla_{\boldsymbol{p}} E(\boldsymbol{p}_\mathcal{M}) \,, \tag{3}$$

where the step size γ is typically chosen to minimize the distance to $\boldsymbol{p}^*$ along the predicted gradient direction (Wagner et al., 2026). Samples whose predicted gradient points away from the ground state are not updated, so that difficult densities remain in the cache until the model learns to improve

its gradients at these points. The central loss component is the gradient-descent-improvement (GDI) loss (Remme & Hamprecht, 2026), which requires that every optimization step with step size λ moves the coefficients closer to the true ground state by at least a predefined contraction factor $0 < \beta < 1$. With $\boldsymbol{p}' = \boldsymbol{p} - \lambda \nabla_{\boldsymbol{p}} E(\boldsymbol{p})$ and $[x]_+ = \max(0, x)$, the loss is

$$\mathcal{L}_{\text{GDI}} = \left[\|\boldsymbol{p}' - \boldsymbol{p}^*\| - \beta \|\boldsymbol{p} - \boldsymbol{p}^*\| \right]_+ . \tag{4}$$

To also enforce the correct energy value at the minimum, the energy-bound loss (Wagner et al., 2026) penalizes predictions that leave empirical quadratic bounds,

$$\mathcal{L}_{\text{EB}} = \left[E^* + a_{\min} \|\boldsymbol{p} - \boldsymbol{p}^*\|^2 - E(\boldsymbol{p}) \right]_+ + \left[E(\boldsymbol{p}) - E^* - a_{\max} \|\boldsymbol{p} - \boldsymbol{p}^*\|^2 \right]_+ . \tag{5}$$

Furthermore, a fraction $0 \leq q < 1$ of each training batch is evaluated at the exact ground-state coefficients and penalized with L_1 losses on the predicted energy, $\mathcal{L}_E = |E(\boldsymbol{p}^*) - E^*|$, and on its gradient, $\mathcal{L}_{\nabla E} = \|\nabla_{\boldsymbol{p}} E(\boldsymbol{p}^*)\|_1$, which should vanish at the desired stationary point.

3.3 Surrogate functionals for molecular properties

In orbital-free DFT, the functional is the total electronic energy, and its minimization corresponds to the physically motivated variational principle. The construction itself, however, can in principle be imposed on any scalar property. Wagner et al. (2026) demonstrated that this formulation shows superior generalization compared to methods that learn the energy directly from molecular geometry. This motivates us to also model other scalar molecular properties P using surrogate functionals $F_P[\boldsymbol{p}]$. For training, we use the same pipeline as described in Section 3.2 and simply replace the ground-state value of the functional with the label of the target property, $E^* \to F_P^*$. In the following, we omit the subscript P whenever the property is clear from context. A schematic overview of this method compared to geometry-only and density-input models is shown in Figure 1. We provide further information on the functional parameterization and the target scaling in Appendix A.4.

3.4 Architecture

In this section, we introduce beSEN, an equivariant message-passing neural network architecture based on eSEN (Fu et al., 2025) and inspired by components of BOA (Klockow et al., 2026). This architecture performs especially well in combination with the surrogate functional framework, as can be seen in Table 3. While the overall structure of the model remains very similar to eSEN, specific components of the message passing, node update and readout are changed. A direct comparison of the two architectures is shown in Figure 2, and the newly introduced components are described in the following subsections.

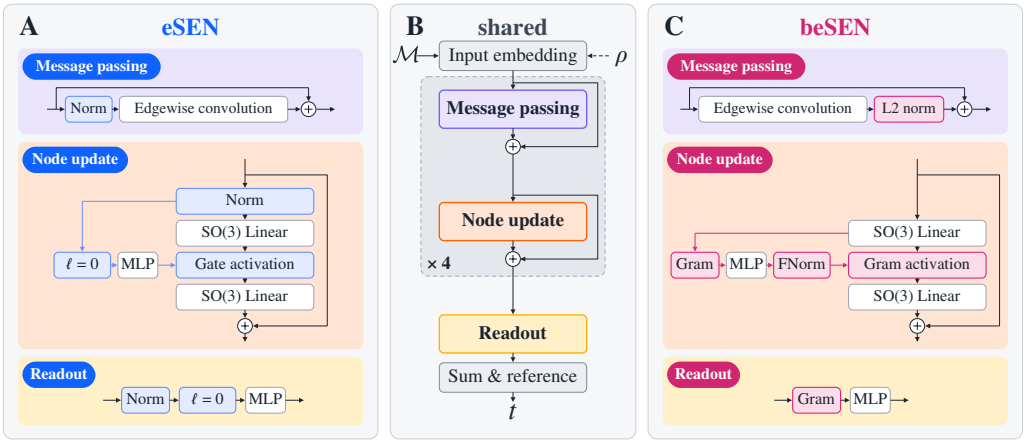

Figure 2: Direct comparison of the eSEN and beSEN architectures. The two architectures share the same overall structure, but differ in multiple details of the message passing, node update and readout. The beSEN architecture is used for both the density prediction and the property prediction models, yielding superior results in both of these settings.

3.4.1 EMBEDDING

The initial embedding is shared by the eSEN and beSEN architectures. It can be split into two parts: the first consumes the molecular geometry and the second consumes the density coefficients. The geometry embedding is the standard eSEN embedding (Fu et al., 2025). The density embedding is a simple equivariant linear layer that maps the density coefficients to the feature dimension of the model, exactly as described in Wagner et al. (2026). The two embeddings are then summed to produce the initial node features.

3.4.2 MESSAGE PASSING

The message passing step differs between the two architectures: eSEN uses a pre-norm residual connection, whereas beSEN uses a post-norm residual connection (Xiong et al., 2020). Additionally, beSEN uses a novel *L2 norm* layer, which computes the mean per-channel L2 norm

$$\mu_n = \mathrm{mean}_a \sqrt{\sum_p x_{apn}^2} \tag{6}$$

for the first batch and then freezes this value to normalize the node features during training and inference. The resulting μ_n is clamped at a minimum of 1×10^{-6}. An additional learned per-channel weight γ_n is used to scale the normalized features:

$$\tilde{x}_{apn} = \gamma_n \frac{x_{apn}}{\mu_n}. \tag{7}$$

This normalization layer is closely related to the normalization used in BOA (Klockow et al., 2026), but without the use of an explicit metric.

3.4.3 NODE UPDATE

The node update of the eSEN and beSEN architectures differ in two places. First, beSEN omits the pre-norm in the node update. Second, beSEN uses a *Gram activation* instead of the standard *Gate activation* (Weiler et al., 2018). The motivation behind both is similar: scalar features are used to gate the tensorial features. In contrast to the *Gate activation*, which normally uses the scalar part of the current node features, the *Gram activation* builds a Gram matrix

$$G_{mn}^a = \sum_p x_{apm} x_{apn} \tag{8}$$

of the feature channels and uses these scalars to gate the tensorial features. A subset of this Gram matrix is then chosen by only considering the first N_{gate} channels along the first dimension, i.e. restricting the matrix to $\tilde{\mathbf{G}}^{\mathbf{a}} \in \mathbb{R}^{N_{\text{gate}} \times N_c}$ with $\tilde{G}_{mn}^a = G_{mn}^a$ for $m \in \{1, \ldots, N_{\text{gate}}\}$ and $n \in \{1, \ldots, N_c\}$, which is flattened and then passed through an MLP to produce the gating matrix

$$\mathbf{W}^a = \mathrm{MLP}\big(\tilde{\mathbf{G}}^a\big), \qquad \mathrm{MLP}\colon \mathbb{R}^{N_{\text{gate}} \times N_c} \to \mathbb{R}^{N_c \times N_c}. \tag{9}$$

This gating matrix $\mathbf{W}^a \in \mathbb{R}^{N_c \times N_c}$ is then passed through a *Frobenius normalization* layer

$$\mu = \mathrm{mean}_a \sqrt{\sum_{mn} (W_{mn}^a)^2}, \qquad \tilde{W}_{mn}^a = \gamma \frac{W_{mn}^a}{\mu}. \tag{10}$$

Similar to the *L2 norm* layer, the resulting μ is clamped at a minimum of 1×10^{-6}, and an additional learned weight γ is used to scale the normalized gating matrix. The final node feature update is then

$$\tilde{x}_{apn} = \sum_m \tilde{W}_{nm}^a x_{apm}. \tag{11}$$

3.4.4 READOUT

The change of the beSEN readout compared to the eSEN readout is similar in spirit to the change in the node update. Instead of taking the scalar features as input for a readout MLP, as in eSEN, beSEN uses the Gram matrix $\mathbf{G}^a$ of the node features as input to the readout MLP.

3.5 SPEED COMPARISON

Table 1: Forward-pass throughput on QM9, beSEN against eSEN at the training batch size of 128, on one idle NVIDIA A100 40GB. Median over 100 distinct full batches drawn from the training split. beSEN has 6.9 M parameters against eSEN's 2.7 M.

Model type	beSEN ms/batch	eSEN ms/batch
Geometry only	**38.4**	44.9
Density input	**39.2**	47.3

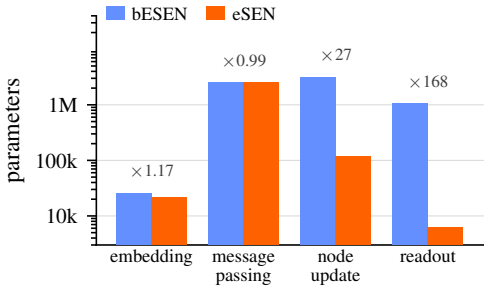

Figure 3: Number of parameters in the eSEN and beSEN architectures. While most of the parameters in the eSEN are in the message passing block, the parameters of beSEN are more evenly distributed between the message passing and the per-node update and readout.

To offer a fair comparison between the eSEN and beSEN architectures on our training task, we measure the forward-pass throughput on the QM9 dataset to make sure that both architectures need a similar amount of time for training and inference. While beSEN has more than twice the number of parameters as the eSEN model used here, it is still slightly faster, as shown in Table 1. Figure 3 shows why the two are not in conflict: the models are the same size in the per-edge message passing, and beSEN's extra parameters sit in the per-node stages, which are far cheaper to evaluate. The throughput is measured on one idle NVIDIA A100 40GB, with a training batch size of 128. The median over 100 distinct full batches drawn from the training split is reported.

3.6 PREDICTING THE GROUND-STATE ELECTRON DENSITY DIRECTLY

Even though the surrogate functional framework in principle allows the simultaneous prediction of the ground-state density and the target property through variational minimization, this process is computationally expensive, requiring ~ 100 model forward calls for the property prediction of one molecule. Predicting the ground-state density coefficients (in the orthonormalized basis) directly with a second neural network and passing them to the main model yields comparable performance. This way, we reduce inference to two model forward calls and at the same time avoid the explicit orthonormalization of the density coefficients. As architecture for the density network, we use the same backbone as described in Section 3.4. For the output head, we use an $\mathrm{SO}(3)$ equivariant MLP that maps to a tensorial feature, which matches the representation of the basis coefficients.

4 RESULTS

We evaluate the performance of our model on the QM9 dataset (Ramakrishnan et al., 2014), which contains 133 885 small organic molecules with up to 9 heavy atoms (C, N, O, F). Additionally, the generalization of the models to larger molecules from the QM40 dataset (Madushanka et al., 2024) is evaluated.

4.1 QM9 DATASET

We train multiple versions of beSEN on the QM9 dataset that differ in the type of input and training mode, but all use the same training, validation and test split as Gasteiger et al. (2020). One model per property is trained on the geometry only, without any density input. A second model per property is trained on the geometry and the ground-state density as input, but without any surrogate functional constraints. A third model per property is trained on the geometry and the ground-state density as input, with the surrogate functional framework applied. All models are trained with the hyperparameters listed in Appendix A.5. The results are shown in Table 2. On U_0, U, H, G, ZPVE, α and C_v the surrogate functional model outperforms the geometry-only and density-input beSEN models by a large margin. For all of these properties but α, the surrogate functional beSEN also outperforms the previous state-of-the-art GotenNet-L model (Aykent & Xia, 2025).

Table 2: Mean absolute error of the predicted properties on the QM9 test set. Values of prior work are reproduced from Aykent & Xia (2025)

	Extensive							*Non-extensive*				
	U_0	U	H	G	ZPVE	α	C_v	HOMO	LUMO	gap	μ	$\langle R^2 \rangle$
Unit	meV	meV	meV	meV	meV	ma_0^3	$\frac{\text{mcal}}{\text{mol K}}$	meV	meV	meV	mD	ma_0^2
Geometry only												
DimeNet++ (Gasteiger et al., 2022)	6.32	6.28	6.53	7.56	1.21	44	23	24.6	19.5	32.6	29.7	331
SphereNet (Liu et al., 2021)	6.26	6.36	6.33	7.78	1.12	46	22	22.8	18.9	31.1	24.5	268
LEFTNet (Du et al., 2023)	6	7	6	7	1.33	48	23	24	18	40	12	109
EQGAT (Le et al., 2022)	25	25	24	23	2.00	53	24	20	16	32	11	382
ET (Thölke & Fabritiis, 2021)	6.15	6.38	6.16	7.62	1.84	59	26	20.3	17.5	36.1	11	33
HDGNN (An et al., 2023)	8.34	8.12	10	11	1.21	46	23	18	16	32	17	342
Geoformer (Wang et al., 2023)	4.43	4.41	4.39	6.13	1.28	40	22	18.4	15.4	33.8	10	28
Equiformer (Liao & Smidt, 2022)	6.59	6.74	6.63	7.63	1.26	46	23	15.4	14.7	30	12	251
SaVeNet-B (Aykent & Xia, 2023)	5.43	5.48	5.43	6.64	1.18	39	23	18.4	16.3	24.8	9.3	58
EquiformerV2 (Liao et al., 2024)	6.17	6.49	6.22	7.57	1.47	47	23	14.4	13.3	29.0	9.9	186
GotenNet-S (Aykent & Xia, 2025)	3.71	3.67	3.70	5.50	1.09	33	20	16.9	13.9	21.2	7.5	29
GotenNet-B (Aykent & Xia, 2025)	3.43	3.49	3.44	5.19	1.09	32	19	15.2	13.0	20.5	7.2	27
GotenNet-L (Aykent & Xia, 2025)	3.37	3.41	3.30	4.98	1.08	**28**	19	**13.4**	**12.2**	**19.8**	**6.7**	**24**
beSEN	5.3	5.4	5.3	6.7	1.1	41	22	17	15	33	8.2	150
Density input												
beSEN	4.9	4.8	4.8	6.0	1.1	39	22	16	15	34	7.9	140
Surrogate												
beSEN	**2.8**	**2.4**	**2.7**	**4.2**	**0.91**	30	**18**	16	15	32	12	480

To evaluate the contribution of the beSEN architecture to the results achieved on the QM9 dataset, we compare the performance of the beSEN and eSEN architectures on the three evaluated training modes: geometry only, density input and surrogate functional. As shown in Table 3, the beSEN architecture outperforms the eSEN architecture on all three training modes, with the difference growing larger for the density input mode and largest for the surrogate functional mode.
We additionally run an extensive ablation study evaluating the impact of different training configurations. The results are discussed in Appendix A.1.

Table 3: Backbone comparison on U_0 showing the mean absolute error (MAE) in meV on the QM9 test split. beSEN outperforms the eSEN architecture on all three training modes.

Arm	eSEN	beSEN
Geometry only	5.6	**5.3**
Density input	5.8	**4.9**
Surrogate	4.2	**2.8**
gain vs geometry only, density input	0.96×	**1.1**×
gain vs geometry only, surrogate	1.3×	**1.9**×

A direct ground state density prediction model using the beSEN architecture is trained on QM9 electron densities. To evaluate the quality of the predicted densities, we compute the normalized mean absolute error

$$\text{NMAE} = \frac{\int \left|\rho^*(\mathbf{r}) - \rho^{\text{pred}}(\mathbf{r})\right| \mathrm{d}\mathbf{r}}{\int \rho^*(\mathbf{r}) \mathrm{d}\mathbf{r}}. \tag{12}$$

An NMAE of 0.032% is achieved on the QM9 test set. The density error is calculated with respect to the density label in the auxiliary basis, i.e. after density fitting (see Appendix A.3). This leads to a larger error compared to the original KS-DFT density. We evaluate the difference between the label before and after density fitting on 100 random QM9 molecules. The mean NMAE is 0.59% and therefore already more than an order of magnitude larger than the error of the direct density prediction model. The integrals necessary for the calculation of the NMAE are computed with `PySCF` (Sun et al., 2018; 2020) on a level-3 integration grid. Additionally, the impact of using the predicted density instead of the label density on the property prediction results is evaluated. The difference in these two densities is small enough that the property prediction results, both in the density input and surrogate setting, are only marginally affected, as shown in Table 4.

Table 4: Mean absolute error of the predicted properties on the QM9 test set, evaluated at the label density and at the density predicted by the direct density prediction model.

Unit	U_0 meV	U meV	H meV	G meV	ZPVE meV	α ma_0^3	C_v $\frac{\mathrm{mcal}}{\mathrm{mol\,K}}$	HOMO meV	LUMO meV	gap meV	μ mD	$\langle R^2 \rangle$ ma_0^2
Density input												
at the label density	4.9	4.8	4.8	6.0	1.1	39	22	16	15	34	7.9	140
at the predicted density	4.9	4.8	4.8	6.0	1.1	39	22	16	15	34	7.9	140
Surrogate												
at the label density	2.7	2.3	2.6	4.0	0.91	30	17	16	15	32	12	430
at the predicted density	2.8	2.4	2.7	4.2	0.91	30	18	16	15	32	12	480

4.2 Generalizing to larger molecules

Table 5: Mean absolute error of the predicted properties on 600 QM40 molecules from 10 to 40 heavy atoms. The dataset is split into bins sorted by the number of heavy atoms, with width of 5 atoms. From each of these bins 100 molecules with atom types H, C, N, O and F that are neutral and with even electron count are selected. Each evaluated model was trained only on QM9. Especially for the extensive properties, using the surrogate functional approach enables better generalization to larger molecules.

	Extensive							*Non-extensive*				
Unit	U_0 meV	U meV	H meV	G meV	ZPVE meV	α ma_0^3	C_v $\frac{\mathrm{mcal}}{\mathrm{mol\,K}}$	HOMO meV	LUMO meV	gap meV	μ mD	$\langle R^2 \rangle$ ma_0^2
Geometry only												
GotenNet-S	310	300	300	220	20	7300	4900	6300	540	—	270	2.7×10^5
GotenNet-B	430	460	450	270	22	7400	5400	4700	520	—	240	—
beSEN	180	180	190	120	150	3000	2600	310	490	1400	**150**	2.3×10^5
Density input												
beSEN	140	130	150	110	150	3100	3100	300	510	**1200**	**150**	2.4×10^5
Surrogate												
beSEN	**57**	**52**	**49**	**53**	**4.9**	**2400**	**170**	**200**	**470**	1300	**150**	$\mathbf{1.5 \times 10^5}$

The capability to generalize to larger molecules is evaluated on 600 molecules from the QM40 dataset (Madushanka et al., 2024), which contains molecules with 10 to 40 heavy atoms. The molecules are sampled from bins sorted by the number of heavy atoms, with a width of 5 heavy atoms. From each bin, 100 neutral molecules containing only the elements H, C, N, O and F and having an even number of electrons are selected. All evaluated models were trained only on QM9. The beSEN model is evaluated in all three training modes. Additionally, the publicly available GotenNet-S and GotenNet-B models (Aykent & Xia, 2025) are evaluated. For the properties missing for the GotenNet models, no checkpoint was available. The results of the evaluation are shown in Table 5. For the extensive properties U_0, U, H, G, ZPVE, α and C_v the surrogate functional model outperforms the geometry-only and density-input beSEN models by a large margin. The intensive properties HOMO and LUMO as well as $\langle R^2 \rangle$ are also predicted best by the beSEN surrogate functional. The HOMO-LUMO gap is predicted best by the density-input beSEN model, while all three beSEN variants perform equally on the dipole moment μ.

To further illustrate the generalization capabilities of the surrogate functional model, Figure 4 shows the absolute error per atom of the predicted U_0 against the number of heavy atoms in the QM40 molecules. The surrogate functional beSEN shows by far the best behavior with respect to molecule size. While all other models show a clear trend of increasing error with increasing molecule size, the surrogate functional model maintains its accuracy far better, leading to much lower median errors for larger molecules. We provide a similar analysis for the remaining properties in Appendix A.2.

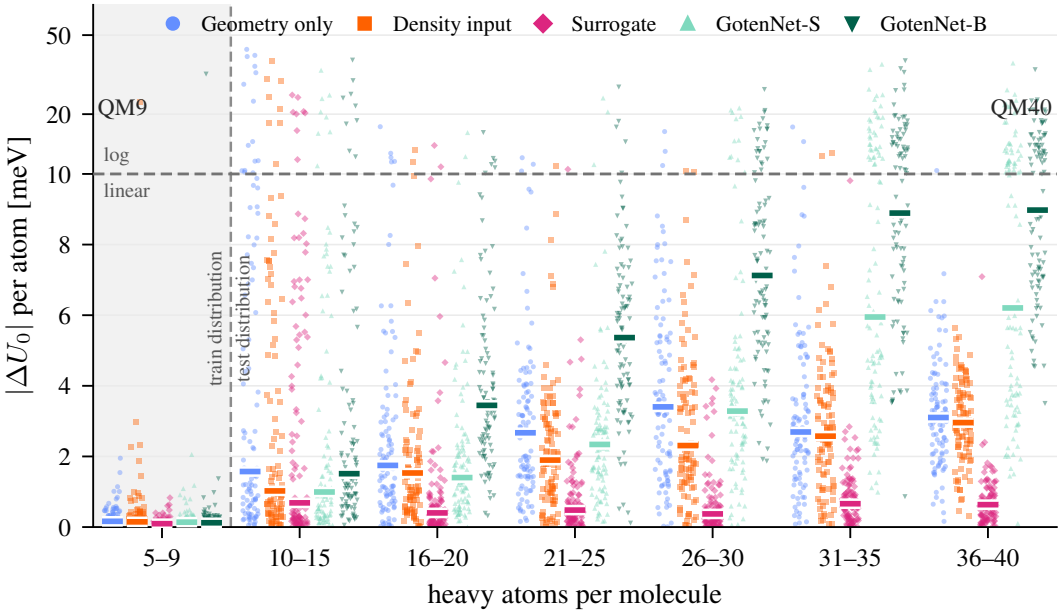

Figure 4: Absolute U_0 error *per atom* against molecule size. The surrogate and the density input models are evaluated on the density as predicted by the density prediction model. The median is shown as a horizontal line.

5 Discussion

In this work, we extend the surrogate functional framework, originally developed in the context of orbital-free density functional theory, to model arbitrary scalar molecular properties. Additionally, a new equivariant message-passing neural network architecture, beSEN, is introduced, which performs especially well in the surrogate functional framework. Employing this framework on the QM9 dataset leads to an improvement on all extensive properties compared to the same architecture trained on geometry-only or geometry and density input. The surrogate functional framework also enables better generalization to larger molecules, as shown on the QM40 dataset.

While the surrogate functional framework has a clear motivation in the context of orbital-free density functional theory, where the electronic energy is governed by a variational principle with respect to the electron density, the motivation for using surrogate functionals for arbitrary molecular properties is less obvious. However, the substantial improvements achieved in this work suggest that the surrogate functional approach is beneficial for all extensive properties. The exact mechanism of this improvement is not yet fully understood and should be investigated in future work.

AI USE STATEMENT

In this work, we used generative AI tools to design or provide feedback on research methodology or experiments and to implement methods. We have *not* used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, or interpret results. Additionally, we used generative AI tools to create or modify scientific figures or images, to create or edit software code, for brainstorming, for sourcing/searching for information, to edit a research paper to improve readability, and to identify relevant literature. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Junyi An, Chao Qu, Zhipeng Zhou, Fenglei Cao, Xu Yinghui, Yuan Qi, and Furao Shen. Hybrid Directional Graph Neural Network for Molecules. October 2023. URL https://openreview.net/forum?id=BBD6KXIGJL.

Dylan M Anstine, Roman Zubatyuk, and Olexandr Isayev. AIMNet2: a neural network potential to meet your neutral, charged, organic, and elemental-organic needs. *Chemical Science*, 16(23): 10228–10244, 2025.

Sarp Aykent and Tian Xia. SaVeNet: A Scalable Vector Network for Enhanced Molecular Representation Learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 42932–42949. Curran Associates, Inc., 2023. doi: 10.52202/075280-1860. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/860c1c657deafe09f64c013c2888bd7b-Abstract-Conference.html.

Sarp Aykent and Tian Xia. GotenNet: Rethinking efficient 3D equivariant graph neural networks. In *The Thirteenth International Conference on Learning Representations*, 2025.

Ilyes Batatia, Dávid Péter Kovács, Gregor N. C. Simm, Christoph Ortner, and Gábor Csányi. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. *Advances in Neural Information Processing Systems*, 35:11423–11436, 2022.

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

Tristan Bereau, Robert A DiStasio, Jr., Alexandre Tkatchenko, and O Anatole von Lilienfeld. Non-covalent interactions across organic and biological subsets of chemical space: Physics-based potentials parametrized from machine learning. *The Journal of Chemical Physics*, 148(24):241706, 2018.

Mihail Bogojeski, Leslie Vogt-Maranto, Mark E Tuckerman, Klaus-Robert Müller, and Kieron Burke. Quantum chemical accuracy from density functional approximations via machine learning. *Nature Communications*, 11(1):5223, 2020.

Mihail Bogojeski, Muhammad R Hasyim, Leslie Vogt-Maranto, Klaus-Robert Müller, Kieron Burke, and Mark E Tuckerman. Enhancing molecular dynamics with equivariant machine-learned densities. *arXiv preprint arXiv:2604.24563*, 2026.

Felix Brockherde, Leslie Vogt, Li Li, Mark E Tuckerman, Kieron Burke, and Klaus-Robert Müller. Bypassing the Kohn–Sham equations with machine learning. *Nature Communications*, 8(1):872, 2017.

weitao Du, Yuanqi Du, Limei Wang, Dieqiao Feng, Guifeng Wang, Shuiwang Ji, Carla Gomes, and Zhi-Ming Ma. A new perspective on building efficient and expressive 3D equivariant graph neural networks. In *Advances in Neural Information Processing Systems*, volume 36, pp. 66647–66674. Curran Associates, Inc., 2023. doi: 10.52202/075280-2910. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d212c6c26603c0eb3c9a6b6432386a1f-Abstract-Conference.html.

Xiang Fu, Brandon M Wood, Luis Barroso-Luque, Daniel S Levine, Meng Gao, Misko Dzamba, and C Lawrence Zitnick. Learning smooth and expressive interatomic potentials for physical property prediction. In *International Conference on Machine Learning*, pp. 17875–17893. PMLR, 2025.

Johannes Gasteiger, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. *arXiv preprint arXiv:2003.03123*, 2020.

Johannes Gasteiger, Shankari Giri, Johannes T. Margraf, and Stephan Günnemann. Fast and Uncertainty-Aware Directional Message Passing for Non-Equilibrium Molecules, April 2022. URL http://arxiv.org/abs/2011.14115. arXiv:2011.14115 [cs.LG].

Andrea Grisafi, Alberto Fabrizio, Benjamin Meyer, David M Wilkins, Clemence Corminboeuf, and Michele Ceriotti. Transferable machine-learning model of the electron density. *ACS central science*, 5(1):57–64, 2019.

P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Physical Review*, 136(3B):B864–B871, 1964.

John A Keith, Valentin Vassilev-Galindo, Bingqing Cheng, Stefan Chmiela, Michael Gastegger, Klaus-Robert Muller, and Alexandre Tkatchenko. Combining machine learning and computational chemistry for predictive insights into chemical systems. *Chemical reviews*, 121(16):9816–9872, 2021.

Manuel Viktor Klockow, Marc Ickler, Peter Lippmann, and Fred A Hamprecht. A function-centric graph neural network approach for predicting electron densities. In *International Conference on Learning Representations*, volume 2026, pp. 69454–69477, 2026.

W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Physical Review*, 140(4A):A1133–A1138, 1965.

Tuan Le, Frank Noé, and Djork-Arné Clevert. Equivariant Graph Attention Networks for Molecular Property Prediction, March 2022. URL http://arxiv.org/abs/2202.09891. arXiv:2202.09891 [cs.LG].

Yi-Lun Liao and Tess Smidt. Equiformer: Equivariant Graph Attention Transformer for 3D Atomistic Graphs. September 2022. URL https://openreview.net/forum?id=KwmPfARgOTD.

Yi-Lun Liao, Brandon M Wood, Abhishek Das, and Tess Smidt. EquiformerV2: Improved equivariant transformer for scaling to higher-degree representations. In *The Twelfth International Conference on Learning Representations*, 2024.

Yi Liu, Limei Wang, Meng Liu, Yuchao Lin, Xuan Zhang, Bora Oztekin, and Shuiwang Ji. Spherical Message Passing for 3D Molecular Graphs. October 2021. URL https://openreview.net/forum?id=givsRXsOt9r.

Per-Olov Löwdin. On the non-orthogonality problem connected with the use of atomic wave functions in the theory of molecules and crystals. *The Journal of Chemical Physics*, 18(3):365–375, 1950.

Ayesh Madushanka, Renaldo T Moura, Jr., and Elfi Kraka. QM40, realistic quantum mechanical dataset for machine learning in molecular science. *Scientific Data*, 11(1):1376, 2024.

Wenhui Mi, Kai Luo, SB Trickey, and Michele Pavanello. Orbital-free density functional theory: An attractive electronic structure method for large-scale first-principles simulations. *Chemical Reviews*, 123(21):12039–12104, 2023.

Felix Musil, Andrea Grisafi, Albert P Bartók, Christoph Ortner, Gábor Csányi, and Michele Ceriotti. Physics-inspired structural representations for molecules and materials. *Chemical Reviews*, 121 (16):9759–9815, 2021.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1):140022, 2014.

Roman Remme and Fred A Hamprecht. Surrogate functionals for machine-learned orbital-free density functional theory. *The Journal of Chemical Physics*, 165(3):034117, 2026.

Roman Remme, Tobias Kaczun, Tim Ebert, Christof A Gehrig, Dominik Geng, Gerrit Gerhartz, Marc K Ickler, Manuel V Klockow, Peter Lippmann, Johannes S Schmidt, Simon Wagner, Andreas Dreuw, and Fred A Hamprecht. Stable and accurate orbital-free density functional theory powered by machine learning. *Journal of the American Chemical Society*, 147(32):28851–28859, 2025.

Kristof T Schütt, Huziel E Sauceda, Pieter-Jan Kindermans, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet – a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24):241722, 2018.

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

Xuecheng Shao, Lukas Paetow, Mark E Tuckerman, and Michele Pavanello. Machine learning electronic structure methods based on the one-electron reduced density matrix. *Nature Communications*, 14(1):6281, 2023.

Qiming Sun, Timothy C Berkelbach, Nick S Blunt, George H Booth, Sheng Guo, Zhendong Li, Junzi Liu, James D McClain, Elvira R Sayfutyarova, Sandeep Sharma, et al. Pyscf: the python-based simulations of chemistry framework. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 8(1):e1340, 2018.

Qiming Sun, Xing Zhang, Samragni Banerjee, Peng Bao, Marc Barbry, Nick S Blunt, Nikolay A Bogdanov, George H Booth, Jia Chen, Zhi-Hao Cui, et al. Recent developments in the pyscf program package. *The Journal of chemical physics*, 153(2), 2020.

Philipp Thölke and Gianni De Fabritiis. Equivariant Transformers for Neural Network based Molecular Potentials. October 2021. URL https://openreview.net/forum?id=zNHzqZ9wrRB.

Ulises A Vergara-Beltran and Juan I Rodríguez. An efficient zero-order evolutionary method for solving the orbital-free density functional theory problem by direct minimization. *J. Chem. Phys.*, 159(12):124102, 2023.

O. Anatole von Lilienfeld, Klaus-Robert Müller, and Alexandre Tkatchenko. Exploring chemical compound space with quantum-based machine learning. *Nature Reviews Chemistry*, 4(7):347–358, 2020.

Simon Wagner, Marc K Ickler, Manuel V Klockow, Fred A Hamprecht, and Roman Remme. Orbital-free surrogate functionals yield transferable interatomic potentials and electron densities. *arXiv preprint arXiv:2609.21882*, 2026.

Yusong Wang, Shaoning Li, Tong Wang, Bin Shao, Nanning Zheng, and Tie-Yan Liu. Geometric Transformer with Interatomic Positional Encoding. In *Advances in Neural Information Processing Systems*, volume 36, pp. 55981–55994. Curran Associates, Inc., 2023. doi: 10.52202/075280-2441. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/aee2f03ecb2b2c1ea55a43946b651cfd-Abstract-Conference.html.

Maurice Weiler, Mario Geiger, Max Welling, Wouter Boomsma, and Taco Cohen. 3d steerable cnns: Learning rotationally equivariant features in volumetric data. *Advances in Neural information processing systems*, 31, 2018.

Dahvyd Wing, Mihail Bogojeski, Szabolcs Goger, Klaus-Robert Müller, and Alexandre Tkatchenko. Accurate and transferable intermolecular potential based on machine-learned molecular electron density. *arXiv preprint arXiv:2608.20753*, 2026.

Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tieyan Liu. On layer normalization in the transformer architecture. In *International conference on machine learning*, pp. 10524–10533. PMLR, 2020.

He Zhang, Siyuan Liu, Jiacheng You, Chang Liu, Shuxin Zheng, Ziheng Lu, Tong Wang, Nanning Zheng, and Bin Shao. Overcoming the barrier of orbital-free density functional theory for molecular systems using deep learning. *Nature Computational Science*, 4(3):210–223, 2024.

A APPENDIX

A.1 ABLATION EXPERIMENTS

Table 6 summarizes ablation experiments on U_0 in which individual components of the surrogate training are removed or modified. All ablated configurations yield a higher error than the full setup, which combines all training components.

Table 6: Ablations of different pieces of the surrogate machinery, on U_0, QM9 test split. *density* determines whether the model receives the predicted ground-state electron density as input. *stat* is the stationarity constraint $\nabla_{\boldsymbol{p}} F[\boldsymbol{p}^*] = 0$ at the ground state, *bound* and *GDI* are the energy-bound and GDI loss respectively and *cache* says whether a cache is used. The minao column indicates how the cache is initialized at each reset, if MINAO is marked the MINAO initialization is used. Otherwise the cache entry is the ground-state density displaced along a random, electron-count-conserving direction, with the L2 displacement radius r drawn log-uniformly (density $\propto 1/r$) over $[0.002, 0.3]$. *gain* denotes the accuracy improvement relative to the geometry-only model.

density	stat	bound	cache	GDI	minao	MAE	gain
						5.3	1.00×
•						4.9	1.09×
•	•					4.9	1.08×
•		•				4.4	1.20×
•	•	•				4.6	1.16×
•		•	•			3.9	1.35×
•	•	•	•			4.9	1.08×
•		•			•	4.6	1.15×
•	•	•			•	4.6	1.15×
•		•	•		•	4.0	1.34×
•	•	•	•		•	5.7	0.93×
•			•	•	•	3.3	1.59×
•	•	•	•	•		**2.8**	1.88×

A.2 SIZE EXTRAPOLATION FOR ALL TARGET PROPERTIES

Figure 5 extends the size analysis of Figure 4 to all twelve target properties. For the geometry-only, density-input, and surrogate beSEN models, it shows the median absolute error per atom on the QM9 test set and on the 600 QM40 molecules, grouped by the number of heavy atoms.

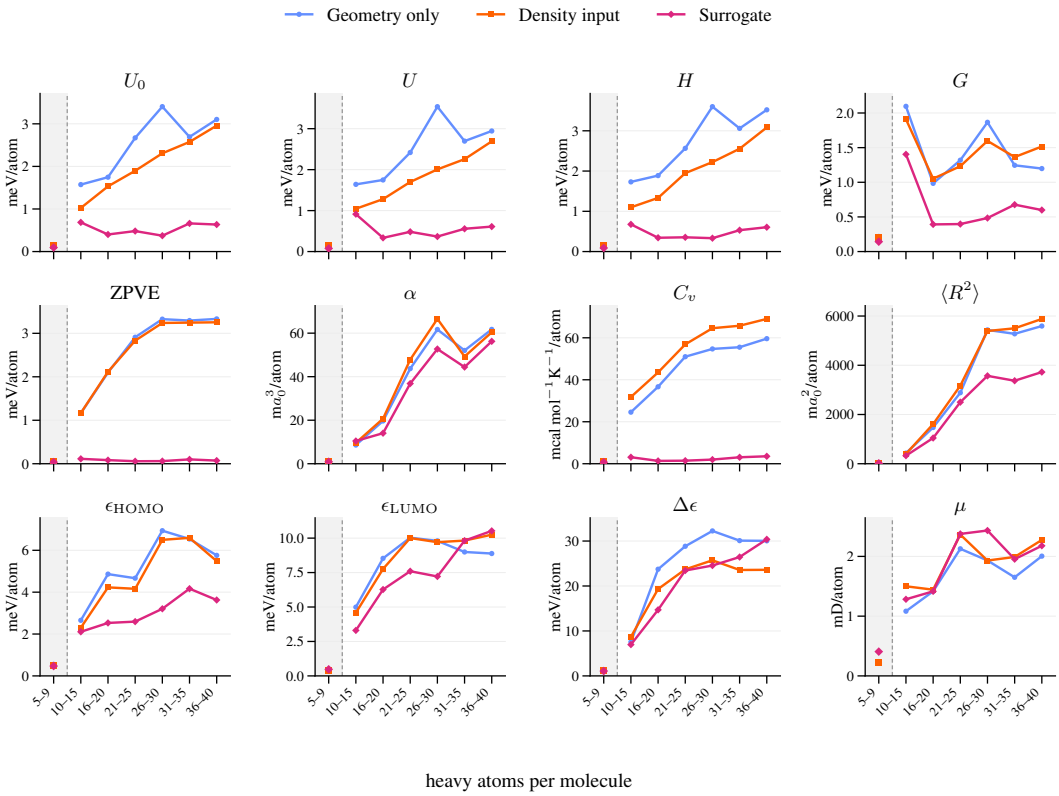

Figure 5: Median absolute error per atom of the geometry-only, density-input, and surrogate beSEN models for all twelve target properties against the number of heavy atoms. The shaded leftmost bin corresponds to the QM9 test set and the remaining bins to the 600 QM40 molecules of Table 5; all models were trained on QM9 only. Errors are divided by the total number of atoms. Error bars and shaded bands indicate 68 % bootstrap confidence intervals of the median.

A.3 GENERATION OF THE ELECTRON DENSITY LABELS

The target properties of all models are taken directly from the QM9 (Ramakrishnan et al., 2014) and QM40 (Madushanka et al., 2024) datasets. The ground-state electron densities are not part of these datasets and are obtained from our own Kohn–Sham DFT calculations. These are carried out with PySCF (Sun et al., 2018; 2020) at the B3LYP/6-31G(2df,p) level of theory that was also used to generate the QM9 dataset.

For the linear density expression in Equation 1 we use an even-tempered Gaussian basis with an exponent ratio of 2.5. The coefficients are obtained from the Kohn–Sham densities via density fitting, using the same fitting objective as in Remme et al. (2025). The fitted coefficients are then orthonormalized as described in Section 3.1.

A.4 Output, and target scaling

A.4.1 Output parametrization

To ensure that the network only needs to predict quantities of order unity for every target, its output is combined with an offset and a scale that are not learned but determined once from the training set. In the following, $\mathbf{h}_a$ denotes the final node features of atom a in molecule $\mathcal{M}$. Furthermore, g_θ denotes the readout network, and $\bar{F}^*$ and σ_{F^*} denote the mean and standard deviation of the training labels F^* (Appendix A.4.2). Depending on the target property, the surrogate functional takes one of the following forms:

- *Extensive properties* (U_0, U, H, G, ZPVE, α, C_v):

$$F[\boldsymbol{p}] = t_0 + \sum_{a\in\mathcal{M}} t_{Z_a} + \frac{1}{c}\sum_{a\in\mathcal{M}} g_\theta(\mathbf{h}_a)\,. \tag{13}$$

  The atomic reference, consisting of element-specific contributions t_Z and a constant term t_0, is obtained by linear least-squares regression of the training targets as in (Zhang et al., 2024). The summed per-atom contributions are divided by a fixed normalization constant c, as in EquiformerV2 (Liao et al., 2024).

- *Intensive properties* (HOMO, LUMO, gap):

$$F[\boldsymbol{p}] = \bar{F}^* + \sigma_{F^*}\, g_\theta\Big(\frac{1}{N_\mathcal{M}}\sum_{a\in\mathcal{M}} \mathbf{h}_a\Big)\,, \tag{14}$$

  where $N_\mathcal{M}$ is the number of atoms. Mean-pooling the node features before the readout ensures that the prediction does not grow with molecular size.

- *Dipole moment μ and electronic spatial extent $\langle R^2\rangle$*: Following GotenNet (Aykent & Xia, 2025), we use the dedicated output heads

$$F_\mu[\boldsymbol{p}] = \sigma_{F^*}\Big\|\sum_{a\in\mathcal{M}}\big(q_a\mathbf{R}_a + \mathbf{d}_a\big)\Big\|\,,\qquad F_{R^2}[\boldsymbol{p}] = \sigma_{F^*}\sum_{a\in\mathcal{M}}\|\mathbf{R}_a - \mathbf{R}_\mathrm{c}\|^2\, x_a\,, \tag{15}$$

  where the atomic charges q_a, atomic dipoles $\mathbf{d}_a$, and scalars x_a are predicted from the node features $\mathbf{h}_a$, $\mathbf{R}_a$ are the atomic positions, and $\mathbf{R}_\mathrm{c}$ is the center of mass.

A.4.2 Target scaling

The hyperparameters of the surrogate losses (Table 9) are absolute in the units of the target: the GDI loss uses a fixed step size λ, which ties the required gradient magnitude, and hence the curvature of the functional, to the units of the target, and the energy-bound loss confines the functional between two fixed parabolas. Since the target properties differ strongly in their units and spread, the same hyperparameters would therefore strike a different balance between fitting the property and shaping the minimum of the functional for every target. To make them transferable, we multiply the labels F_P^* of each target property P by the factor

$$k_P = \frac{\sigma_{U_0}}{\sigma_{F^*}}\,, \tag{16}$$

For extensive properties $\sigma_{F_P^*}$ is computed after subtracting the atomic reference (Appendix A.4.1). All quantities in Appendix A.4.1 are computed from the scaled labels.

A.5 HYPERPARAMETERS

Tables 7–9 list the most relevant hyperparameters. All property models (geometry-only, density-input, and surrogate) as well as the direct density prediction model share the architecture and training settings of Tables 7 and 8; the settings of Table 9 apply to the surrogate models only.

Table 7: Architecture hyperparameters of beSEN and of the eSEN baseline.

	beSEN	eSEN
Number of layers	4	4
Channels	32	32
Maximum degree l_{max}	4	4
Cutoff radius (Bohr)	6.0	6.0
Gaussian radial basis functions	32	128
Edge channels	32	32
Hidden channels of the message MLP	64	64
Residual connection	post-norm	pre-norm
Normalization	L2 norm	RMS norm
Activation	Gram	gate
Gram activation: gating channels N_{gate}	16	–
Gram activation: hidden channels of the gating MLP	512	–
Hidden channels of the readout MLP	1024, 512	64, 64
LayerScale initialization	0.1	–
Number of parameters	6.9 M	2.7 M

Table 8: Training hyperparameters, shared by all models.

Optimizer	AdamW, $(\beta_1, \beta_2) = (0.95, 0.99)$
Learning rate	10^{-3}, cosine annealing to 0
Epochs	500
Batch size	128
Loss function	L_1 (property / density targets)

Table 9: Hyperparameters of the surrogate training (Section 3.2). The energy bounds refer to the scaled labels (Appendix A.4.2).

GDI step size λ	0.04
GDI contraction factor β	0.86
Energy-bound factors a_{min}, a_{max}	2, 5
Fraction q of ground-state samples	0.5
Loss weights of $\mathcal{L}_{\text{GDI}}$, $\mathcal{L}_{\text{EB}}$, $\mathcal{L}_E$, $\mathcal{L}_{\nabla E}$	20, 1, 1, 1
Cache step size γ	adaptive, minimum 10^{-4}
Cache initialization	MINAO initial guess
Cache reset probability per update	0.02