A. Quantum-Chemical Calculations

For the calculation of the energy E we use the fast semi-empirical Parametrized Method 6 (PM6) (Stewart, 2007). In particular, we use the implementation from the software package SPARROW (Husch et al., 2018; Bosia et al., 2019). For each calculation, a molecular charge of zero and the lowest possible spin multiplicity are chosen. All calculations are spin-unrestricted.

Limitations of semi-empirical methods are highlighted in, for example, recent work by Husch & Reiher (2018). More accurate methods such as approximate density functionals need to be employed especially for systems containing transition metals.

For the quantum-chemical calculations to converge reliably, we ensured that atoms are not placed too close (< 0.6 Å) nor too far away from each other (> 2.0 Å). If the agent places an atom outside these boundaries, the minimum reward of -0.6 is awarded and the episode terminates.

B. Learning the Dihedral Angle

We experimentally validate the benefits of learning $|\psi| \in [0, \pi]$ and $\kappa \in \{-1, 1\}$ instead of $\psi \in [-\pi, \pi]$ by comparing the two models on the *single-bag* task with bag CH_4 (methane). Methane is one of the simplest molecules that requires the model to learn a dihedral angle. As shown in Fig. 9, learning the sign of the dihedral angle separately (with κ) speeds up learning significantly. In fact, the ablated model (without κ) fails to converge to the optimal return even after 100 000 steps (not shown).

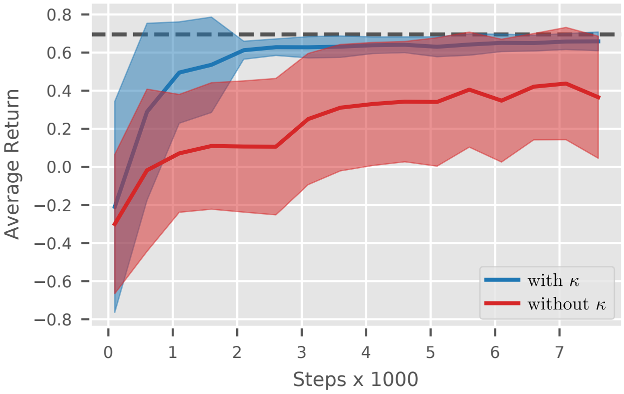

Figure 9. Average offline performance on the *single-bag* task for the bag CH_4 across 10 seeds. Estimating κ and $|\psi|$ separately (with κ) significantly speeds up learning compared to estimating ψ directly (without κ). Error bars show two standard deviations. The dashed line denotes the optimal return.

C. Experimental Details

C.1. Model Architecture

The model architecture is summarized in Table 3. We initialize the biases of each MLP with 0 and each weight matrix as a (semi-)orthogonal matrix. After each hidden layer, a ReLU non-linearity is used. The output activations are shown in Table 3. As explained in the main text, both MLP_f and MLP_e use a masked softmax activation function to guarantee that only valid actions are chosen. Further, we rescale the continuous actions $(\mu_d, \mu_\alpha, \mu_\psi) \in [-1, 1]^3$ predicted by MLP_cont to ensure that $\mu_d \in [d_\text{min}, d_\text{max}]$, $\mu_\alpha \in [0, \pi]$ and $\mu_\psi \in [0, \pi]$. For more details on the SchNet, see the original work (Schütt et al., 2018b).

Table 3. Model architecture for actor and critic networks.

Operation	Dimensionality	Activation
SchNet	$n(\mathcal{C}) \times 4, *, n(\mathcal{C}) \times 64$	$*$ (cf. Table 7)
MLP_β	$e_\text{max}, 128, 32$	linear
tile	$32, n(\mathcal{C}) \times 32$	—
concat	$n(\mathcal{C}) \times (64, 32), n(\mathcal{C}) \times 96$	—
MLP_f	$n(\mathcal{C}) \times 96, n(\mathcal{C}) \times 128, n(\mathcal{C}) \times 1$	softmax
select	$n(\mathcal{C}) \times 96, 96$	—
MLP_e	$96, 128, e_\text{max}$	softmax
concat	$(96, e_\text{max}), 96 + e_\text{max}$	—
MLP_cont	$96 + e_\text{max}, 128, 3$	tanh
MLP_κ	$2 \times 96, 2 \times 128, 2 \times 1$	softmax
pooling	$n(\mathcal{C}) \times 96, 96$	—
MLP_ϕ	$96, 128, 128, 1$	linear

C.2. Hyperparameters

We manually performed an initial hyperparameter search on a single holdout validation seed. The considered hyperparameters and the selected values are listed in Table 4 (*single-bag*), Table 5 (*multi-bag*) and Table 6 (*solvation*). The hyperparameters used for SchNet are shown in Table 7.

D. Baselines

Below, we report how the baselines for the *single-bag* and *multi-bag* tasks were derived. First, we took all molecular structures for a given chemical formula (i.e. bag) from the QM9 dataset (Ruddigkeit et al., 2012; Ramakrishnan et al., 2014). Subsequently, we performed a structure optimization using the PM6 method (as described in Section A) on the structures. This was necessary as the structures in this dataset were optimized with a different quantum-chemical method. Then, the most stable structure was selected and considered *optimal* for this chemical formula; the remaining structures were discarded. Since the undiscounted return is path independent, we determined the return $R(s)$ by com-

Table 4. Hyperparameters for the *single-bag* task. Adapted values for the scalability (large) experiment are in parentheses.

Hyperparameter	Search set	Value (large)
Range $[d_{\text{min}}, d_{\text{max}}]$ (Å)	—	$[0.95, 1.80]$
Max. atomic number e_{max}	—	10
Workers	—	16
Clipping ϵ	—	0.2
Gradient clipping	—	0.5
GAE parameter λ	—	0.95
VF coefficient c_1	—	1
Entropy coefficient c_2	$\{0.00, 0.01, 0.03\}$	0.01
Training epochs	$\{5, 10\}$	5
Adam stepsize	$\{10^{-4}, 3 \times 10^{-4}\}$	3×10^{-4}
Discount γ	$\{0.99, 1.00\}$	0.99
Time horizon T	$\{192, 256\}$	192 (256)
Minibatch size	$\{24, 32\}$	24 (32)

Table 5. Hyperparameters for the *multi-bag* task.

Hyperparameter	Search set	Value
Range $[d_{\text{min}}, d_{\text{max}}]$ (Å)	—	$[0.95, 1.80]$
Max. atomic number e_{max}	—	10
Workers	—	16
Clipping ϵ	—	0.2
Gradient clipping	—	0.5
GAE parameter λ	—	0.95
VF coefficient c_1	—	1
Entropy coefficient c_2	$\{0.00, 0.01, 0.03\}$	0.01
Training epochs	$\{5, 10\}$	5
Adam stepsize	$\{10^{-4}, 3 \times 10^{-4}\}$	3×10^{-4}
Discount γ	$\{0.99, 1.00\}$	0.99
Time horizon T	$\{384, 512\}$	384
Minibatch size	$\{48, 64\}$	48

puting the total interaction energy in the canvas $\mathcal{C}$, i.e.

$$R(s) = E(\mathcal{C}) - \sum_{i=1}^{N} E(e_i), \tag{9}$$

where N is the number of atoms placed on the canvas.

The baseline for the *solvation* task was determined in the following way. 12 molecular clusters were generated by randomly placing n H_2O molecules around the solute molecule (in the main text $n = 5$). Subsequently, the structure of these clusters was optimized with the PM6 method (as described in Section A). Similar to Eq. (9), the undiscounted return of each cluster can be computed:

$$R(s) = E(\mathcal{C}) - E(\mathcal{C}_0) - \sum_{i=1}^{N} \{E(e_i) + \rho \|x_i\|_2\}, \tag{10}$$

where the distance penalty $\rho = 0.01$. Finally, the maximum return over the optimized clusters was determined.

Table 6. Hyperparameters for the *solvation* task.

Hyperparameter	Search set	Value
Range $[d_{\text{min}}, d_{\text{max}}]$ (Å)	—	$[0.90, 2.80]$
Max. atomic number e_{max}	—	10
Distance penalty ρ	—	0.01
Workers	—	16
Clipping ϵ	—	0.2
Gradient clipping	—	0.5
GAE parameter λ	—	0.95
VF coefficient c_1	—	1
Entropy coefficient c_2	$\{0.00, 0.01, 0.03\}$	0.01
Training epochs	$\{5, 10\}$	5
Adam stepsize	$\{10^{-4}, 3 \times 10^{-4}\}$	3×10^{-4}
Discount γ	$\{0.99, 1.00\}$	0.99
Time horizon T	$\{384, 512\}$	384
Minibatch size	$\{48, 64\}$	48

Table 7. Hyperparameters for SchNet (Schütt et al., 2018a) used in all experiments.

Hyperparameter	Search set	Value
Number of interactions	—	3
Cutoff distance (Å)	—	5.0
Number of filters	—	128
Number of atomic basis functions	$\{32, 64, 128\}$	64

E. Additional Results

E.1. Single-bag Task

In Fig. 10, we show a selection of molecular structures generated by trained models for the bags C_4H_7N and C_3H_8O. Further, since the agent is agnostic to the concept of molecular bonds, it is able to build multiple molecules if it results in a higher return. An example of a bimolecular structure generated by a trained model for the bag C_3H_8O is shown in Fig. 11. Finally, in Fig. 12, we showcase a set of generated molecular structures that are not chemically valid.

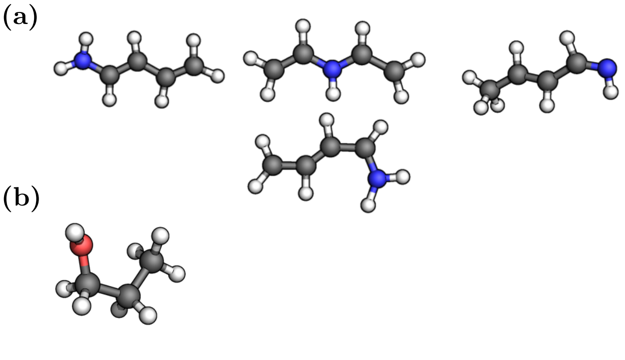

Figure 10. Selection of molecular structures generated by trained models for the bags C_4H_7N (a) and C_3H_8O (b).

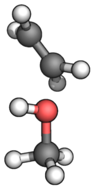

Figure 11. Bimolecular structure generated by a trained model for the bag C_3H_8O in the *single-bag* task.

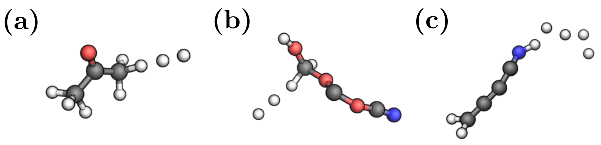

Figure 12. Selection of chemically invalid molecular structures generated by trained models for the bags C_3H_8O (a), $C_3H_5NO_3$ (b), and C_4H_7N (c).

E.2. Solvation Task

In Fig. 13, we report the average offline performances of agents placing 5 H_2O molecules around the solutes (i.e, $\mathcal{C}_0$) acetonitrile and ethanol. As can be seen, the agents are able to accurately place water molecules such that they interact with the solute. However, we stress that more accurate quantum-chemical methods for computing the reward are required to describe hydrogen bonds to chemical accuracy.

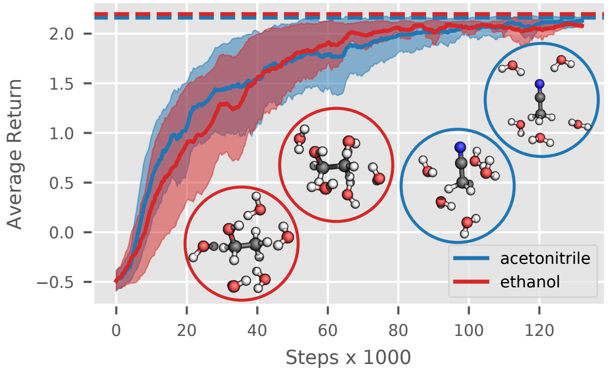

Figure 13. Average offline performances across 10 seeds on the *solvation* task with $n = 5$ and the initial states being acetonitrile and ethanol. Error bars show two standard errors. The plot is smoothed across five evaluations for better readability. The dashed lines denote the optimal returns. A selection of molecular clusters generated by trained models are shown in circles.

In Fig. 14, we compare the average offline performance of two agents placing in total 10 H and 5 O atoms around a formaldehyde molecule. One agent is given 5 H_2O bags consecutively following the protocol of the *solvation* task as described in the main text, another is given a single $H_{10}O_5$ bag. Their average offline performances are shown in Fig. 14 in blue and red, respectively. It can be seen that giving the agent 5 H_2O bags one at a time instead of a single $H_{10}O_5$ bag improves performance.

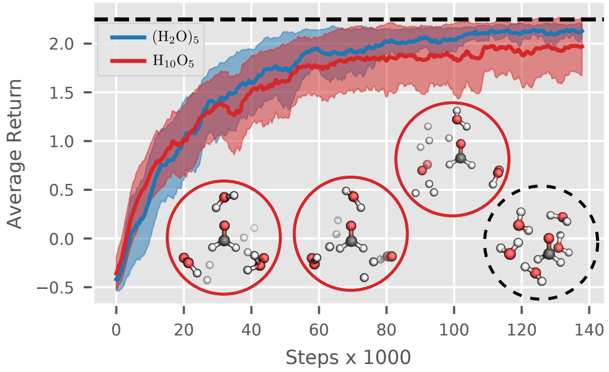

Figure 14. Average offline performance for the *solvation* task with $n = 5$ (blue) and placing atoms from a single $H_{10}O_5$ bag (red). In both experiments, $\mathcal{C}_0$ is formaldehyde. Error bars show two standard errors. The plot is smoothed across five evaluations for better readability. The dashed line denotes the optimal return. A selection of molecular clusters generated by models trained on the $H_{10}O_5$ bag are shown in red solid circles; for comparison, a stable configuration obtained through structure optimization is depicted in a black dashed circle.

E.3. Generalization and Transfer Learning

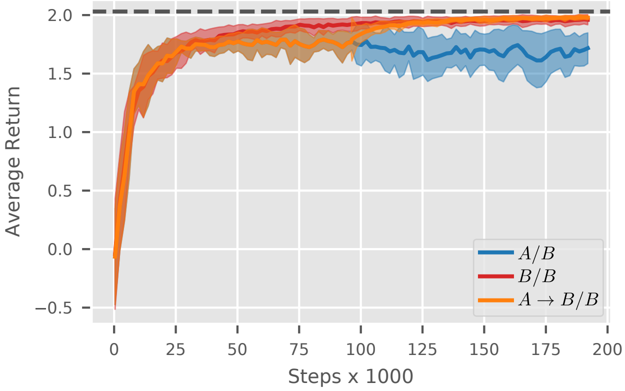

Figure 15. Average offline performance for agents A/B: trained on bags A of size 6 and tested on bags B of size 8, B/B: trained and tested on B, and $A \to B/B$: trained on A for 96 000 steps, and fine-tune and tested on B. See main text for more details. Error bars show two standard deviations. The dashed line denotes the optimal average return.

To assess the generalization capabilities of our agent when faced with previously unseen bags, we train an agent on bags $A = \{C_2H_2O_2, C_2H_3N, C_3H_2O, C_3N_2O, CH_3NO, CH_4O\}$ of size 6 and test on bags $B = \{C_3H_2O_3, C_3H_4O, C_4H_2O_2, CH_4N_2O, C_4N_2O_2, C_5H_2O\}$ of size 8. As shown in Fig. 15, the agent A/B achieves an average return of 1.79, which is approximately 88% of the optimal return. In comparison, an agent trained and tested on B (B/B) reaches an average return of 1.96 (or 0.97% of the optimal return). We additionally train an agent on A for $96\,000$ steps, and then fine-tune and test on B. The agent $A \rightarrow B/B$ reaches the same performance as if trained from scratch within $20\,000$ steps of fine-tuning, showing successful transfer. We anticipate that training on more bags and incorporating best practices from multi-task learning would further improve performance.