

Electronic Supplementary Information

Embedded machine-readable molecular representation for resource-efficient deep learning applications

Emilio Nuñez-Andrade^{*a}, Isaac Vidal-Daza^{a,b}, James W. Ryan^{a,c}, Rafael Gómez-Bombarelli^d and Francisco J. Martín-Martínez^{*e,a}

Contents

1 Methods	1
1.1 Embedding One Hot Encoding	1
1.2 VAE model architecture	2
1.3 RNN model architecture	2
2 Results and discussion	2
2.1 Preliminary considerations about eOHE	2
2.2 Complementary Memory Results	3
3 Supplementary Figures and Tables	4

List of Figures

1	(a) Values for $r(q = 8, k)$ given by Equation 5. Check ESI-Table 1 for summary of r values. (b) Plots of Equation 4 and Equation 6 with $q = 8$ and $r(q = 8, k)$. v_1 scales linearly and v_2 scales as a normalized power of 2.	2
2	VAE model metric results for QM9 database subsets.	5
3	VAE model metric results for QM9 database subsets (continue).	6
4	VAE model metric results for GDB-13 database subsets.	7
5	VAE model metric results for GDB-13 database subsets (continue).	8
6	VAE model metric results for ZINC database subsets.	9
7	VAE model metric results for ZINC database subsets.	10
8	RNN model metric results for QM9 database subsets.	11
9	RNN model metric results for QM9 database subsets (continue).	12
10	RNN model metric results for GDB-13 database subsets.	13
11	RNN model metric results for GDB-13 database subsets (continue).	14
12	RNN model metric results for ZINC database subsets.	15

13	RNN model metric results for ZINC database subsets (continue).	16
----	--	----

List of Tables

1	Values calculated for r , labeled as A to H in the colorbars of Figure 1(b), (c), this values are calculated with Equation 4 and Equation 6 for $v_1(r, q)$ and $v_2(r, q)$, respectively.	2
2	Summary of empty classes added to the dictionary of tokens for the VAE model trained with eOHE.	3
3	Summary of empty classes added to the dictionary of tokens for the RNN model trained with eOHE.	3
4	Differences in the use of RAM memory (in GB) between the VAE model trained with OHE and the same model trained with eOHE-v1 (columns V1) and eOHE-v2 (columns V2), across various database subsets. Each data point is calculated using the mean value of ten independent replicates. Colors indicate the molecular representation used: ■ SMILES, ■ DeepSMILES and ■ SELFIES.	3
5	Differences in the use of RAM memory (in GB) between the RNN model trained with OHE and the same model trained with eOHE-v1 (columns V1) and eOHE-v2 (columns V2), across various database subsets. Each data point is calculated using the mean value of ten independent replicates. Colors indicate the molecular representation used: ■ SMILES, ■ DeepSMILES and ■ SELFIES.	4
6	Values ℓ , p , q and m for eOHE used during training of VAE model for subsets of QM9 database.	17
7	Values ℓ , p , q and m for eOHE used during training of RNN model for subsets of QM9 database.	17
8	Values ℓ , p , q and m for eOHE used during training of VAE model for subsets of GDB-13 database.	18
9	Values ℓ , p , q and m for eOHE used during training of RNN model for subsets of GDB-13 database.	18
10	Values ℓ , p , q and m for eOHE used during training of VAE model for subsets of ZINC database.	19
11	Values ℓ , p , q and m for eOHE used during training of RNN model for subsets of ZINC database.	19

1 Methods

1.1 Embedding One Hot Encoding

ESI-Figure 1(a) shows the values of r for all the k -indexes in the dictionary of tokens ($\ell = 24$) for 4-nitro-1H-pyrrol-2-ol, in

^a Department of Chemistry, Swansea University, Singleton Park, Sketty, SA28PP, Swansea, UK. E-mail: 2132253@swansea.ac.uk

^b Grupo de Modelización y Diseño Molecular, Departamento de Química Orgánica, Facultad de Ciencias, Universidad de Granada, 18071 Granada, Spain.

^c Centre for Integrative Semiconductor Materials (CISM), Swansea University, Swansea SA1 8EN

^d Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, United States of America.

^e Department of Chemistry, Faculty of Natural, Mathematical, and Engineering Sciences, King's College London, SE1 1DB, London, UK. E-mail: francisco.martin-martinez@kcl.ac.uk

accordance with Equation 5, while ESI-Figure 1(b) shows the embedded values, $v_1(r, q)$ and $v_2(r, q)$, resulting from applying the eOHE-v1 method with Equation 4, or the eOHE-v2 method with Equation 6 and $q = 8$ for 4-nitro-1H-pyrrol-2-ol. While eOHE-v1 scales linearly from 0 to 1, eOHE-v2 scales as an exponentially normalized function by a power of 2. ESI-Table 1 summarizes the values for $r(q = 8, k)$, $v_1(r, q = 8)$ and $v_2(r, q = 8)$.

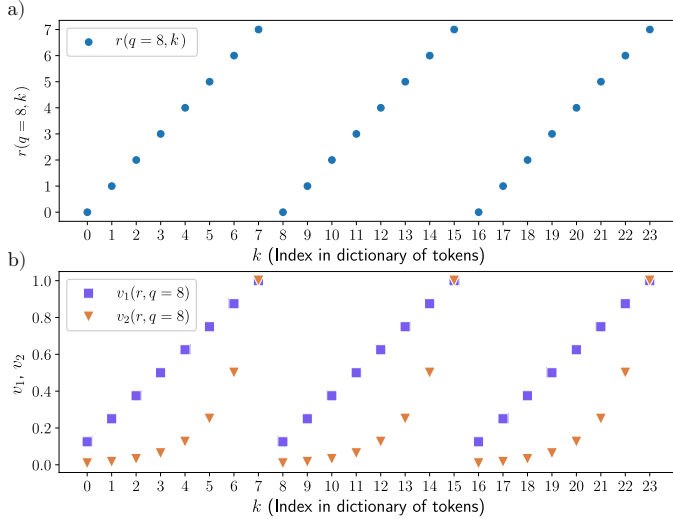


Figure S1 (a) Values for $r(q=8, k)$ given by Equation 5. Check ESI-Table 1 for summary of r values. (b) Plots of Equation 4 and Equation 6 with $q = 8$ and $r(q=8, k)$. v_1 scales linearly and v_2 scales as a normalized power of 2.

Table S1 Values calculated for r , labeled as A to H in the colorbars of Figure 1(b), (c), this values are calculated with Equation 4 and Equation 6 for $v_1(r, q)$ and $v_2(r, q)$, respectively.

k	r	Label	$v_1(r, q)$	$v_2(r, q)$
0, 8, 16	0	A	0.125	0.0078125
1, 9, 17	1	B	0.25	0.015625
2, 10, 18	2	C	0.375	0.03125
3, 11, 19	3	D	0.5	0.0625
4, 12, 20	4	E	0.625	0.125
5, 13, 21	5	F	0.75	0.25
6, 14, 22	6	G	0.875	0.5
7, 15, 23	7	H	1.0	1.0

1.2 VAE model architecture

The VAE model (Krenn *et al.*)¹ consists of two key components: the encoder and the decoder. In the encoder, the model employs 3 densely connected layers with 100, 50, and 25 neurons per layer, respectively. A ReLU activation function follows each linear layer. These linear layers aim to map the input to a 25-dimensional latent space. The decoder consists of a single-layer Gated Recurrent Unit (GRU) with hidden size of 100 neurons. Subsequently, a fully connected linear layer is applied to generate the final output, the number of neurons in the linear layer of the decoder matches the length of the dictionary of

tokens. The learning rate for both the encoder and decoder is 0.002427, the optimizer used was Adam, with a batch size of 190 and 50-50% splitting of data for training and validation, and the loss function was Categorical Cross Entropy. This model monitors the training process, if the model has not improved the reconstruction rate in 20 epochs, the training stops.

1.3 RNN model architecture

The RNN model used (Skinnider *et al.*)² consists of a GRU layer with a hidden size of 512 neurons and 3 layers. After the GRU layers, the model uses a linear fully connected layer for decoding. This decoder layer maps the RNN's output to a space that matches the size of the length of the dictionary of tokens. The learning rate for the model was 0.001, and the batch size was 128. The Adam optimizer was also implemented, and a 90-10% splitting of data was used for training and validation, respectively. The loss function used to train this model was also Categorical Cross Entropy. This model has functioned to monitor the training process and to stop if the model shows evidence of overfitting in the validation data.

2 Results and discussion

2.1 Preliminary considerations about eOHE

Before proceeding with discussing the performance of implementing eOHE into a VAE or RNN model training, there are two practical aspects to consider. First, the impact of adding empty classes to the dictionary of tokens, which is required to achieve factorizable p - q pairs in the training of the models, as mentioned before. Extensive studies indicate that the addition of empty classes, could generate data imbalances that form gradients and weights that create model biases towards majority classes^{3,4}.

The second practical aspect to consider is the dependency of learning outcomes on the order in which the tokens are organized within the dictionary, which is a question that also applies to OHE. We performed several trials with different organization of tokens in the dictionary. We considered four different arrangements: 1) a distribution from the most frequent token to the least one; 2) a distribution with tokens grouped by index, with odd-index tokens on one side of the dictionary, and even-index ones on the other side; 3) a distribution with most frequent tokens in the center of the dictionary and the others arranged by decreasing frequency order towards the sides; and 4) in a randomly distributed arrangement. None of the arrangements had an impact in the performance metrics with eOHE.

Most empty classes were added during trials with the VAE model, although some of them were also included with the RNN model. In the case of the RNN two empty classes in a single dictionary of tokens were sometimes needed. ESI-Table 2 shows the subsets and the encoding representations that required the addition of one empty class to the dictionary during training of the VAE model, and ESI-Table 3 shows the subsets and the encoding representations that required the addition of one or two empty classes to the dictionary during training of the RNN model.

Table S2 Summary of empty classes added to the dictionary of tokens for the VAE model trained with eOHE. White color indicates that no empty class was added to the subset (□ None) green indicates that one class was added when SMILES were encoded (■ SMILES), blue indicates that one empty class was added when DeepSMILES were encoded (■ DeepSMILES), and orange indicates that one empty class was added when SELFIES were encoded (■ SELFIES). White boxes with an × inside (⊗) highlighted those subsets that were not included in the training process.

		No. of molecules ($\times 10^3$)													
		Idx	1	2.5	5	7.5	10	25	50	75	100	125	250	500	
Database Name	QM9	1													
		2													
		3													
		4													
		5													
		6													
		7													
		8													
		9													
		10													
	GDB	1													
		2													
		3													
		4													
		5													
		6													
		7													
		8													
		9													
		10													
	ZINC	1													
		2													
		3													
		4													
		5													
		6													
		7													
		8													
		9													
		10													

Table S3 Summary of empty classes added to the dictionary of tokens for the RNN model trained with eOHE codification. Indigo color indicates that no empty class was added to the subset (□ None) green indicates that one class was added when SMILES were encoded (■ SMILES), blue indicates that one empty class was added when DeepSMILES were encoded (■ DeepSMILES), and orange indicates that one empty class was added when SELFIES were encoded (■ SELFIES). A dot inside a box indicates that two empty classes were needed in that specific case. White boxes with an × inside (⊗) highlighted those subsets that were not included in the training process.

		No. of molecules ($\times 10^3$)													
		Idx	1	2.5	5	7.5	10	25	50	75	100	125	250	500	
Database Name	QM9	1												×	×
		2												×	×
		3												×	×
		4												×	×
		5												×	×
		6												×	×
		7												×	×
		8												×	×
		9												×	×
		10												×	×
	GDB	1											×		
		2											×		
		3											×		
		4											×		
		5											×		
		6											×		
		7											×		
		8											×		
		9											×		
		10											×		
	ZINC	1											×		
		2											×		
		3											×		
		4											×		
		5											×		
		6											×		
		7											×		
		8											×		
		9											×		
		10											×		

The cases that required two empty classes are indicated with a colored cell and a dot inside the box. If only one empty class was required, it was added as token '< 0 >'. If two of them were needed, they were added as tokens '< 0 >' and '< 1 >'. For a complete list of p , q , ℓ values, together with the number of empty classes, m , added to every database subset, model and molecular string representation, check ESI-Table 6 to ESI-Table 11.

Despite the existing literature about data imbalances, our empirical results indicate that any imbalances occurring from adding empty classes to the dictionary of tokens do not affect the performance of eOHE, and that all metrics being considered in our benchmark achieve similar values to those with OHE, which does not have empty classes in its implementation.

2.2 Complementary Memory Results

ESI-Table 4 and ESI-Table 5 show the differences in RAM memory usage by a single NVIDIA A100 Tensor Core GPU for the VAE and RNN models, respectively, when trained using OHE or eOHE. Positive values in the ESI-Table 4 and ESI-Table 5 indicate that OHE required more RAM memory than eOHE, and negative values indicate that OHE required less RAM memory than eOHE.

Table S4 Differences in the use of RAM memory (in GB) between the VAE model trained with OHE and the same model trained with eOHE-v1 (columns V1) and eOHE-v2 (columns V2), across various database subsets. Each data point is calculated using the mean value of ten independent replicates. Colors indicate the molecular representation used: ■ SMILES, ■ DeepSMILES and ■ SELFIES.

		Idx	V1	V2	V1	V2	V1	V2
QM9	No. of molecules ($\times 10^3$)	1	0.004	0.004	0.004	0.004	0.004	0.004
		2.5	0.012	0.012	0.011	0.011	0.012	0.012
		5	0.025	0.025	0.023	0.023	0.025	0.025
		7.5	0.036	0.036	0.035	0.035	0.035	0.035
		10	0.049	0.049	0.045	0.045	0.045	0.045
		25	0.105	0.105	0.103	0.103	0.103	0.103
		50	0.191	0.191	0.186	0.186	0.186	0.186
		75	0.277	0.277	0.271	0.271	0.271	0.271
		100	0.361	0.361	0.350	0.350	0.350	0.350
		125	0.450	0.450	0.434	0.434	0.434	0.434
GDB	No. of molecules ($\times 10^3$)	1	0.005	0.005	0.006	0.006	0.005	0.005
		2.5	0.014	0.014	0.015	0.015	0.013	0.013
		5	0.031	0.031	0.033	0.033	0.026	0.026
		7.5	0.044	0.044	0.046	0.046	0.038	0.038
		10	0.061	0.061	0.063	0.063	0.051	0.051
		25	0.143	0.143	0.158	0.158	0.120	0.120
		50	0.265	0.265	0.299	0.299	0.228	0.228
		75	0.383	0.383	0.434	0.434	0.337	0.337
		100	0.503	0.503	0.570	0.570	0.455	0.455
		250	1.219	1.219	1.387	1.387	1.110	1.110
ZINC	No. of molecules ($\times 10^3$)	500	2.410	2.410	2.744	2.744	2.265	2.265
		1	0.015	0.015	0.019	0.019	0.018	0.018
		2.5	0.039	0.039	0.050	0.050	0.041	0.041
		5	0.087	0.087	0.115	0.115	0.090	0.090
		7.5	0.120	0.120	0.168	0.168	0.127	0.127
		10	0.166	0.166	0.246	0.246	0.172	0.172
		25	0.412	0.412	0.537	0.537	0.406	0.406
		50	0.813	0.813	1.156	1.156	0.798	0.798
		75	1.221	1.221	1.820	1.820	1.199	1.199
		100	1.650	1.650	2.484	2.484	1.692	1.692
		250	4.272	4.272	6.670	6.670	4.338	4.338
		500	8.899	8.899	14.406	14.406	9.225	9.225

Table S5 Differences in the use of RAM memory (in GB) between the RNN model trained with OHE and the same model trained with eOHE-v1 (columns V1) and eOHE-v2 (columns V2), across various database subsets. Each data point is calculated using the mean value of ten independent replicates. Colors indicate the molecular representation used: ■ SMILES, ■ DeepSMILES and ■ SELFIES.

	Idx	V1	V2	V1	V2	V1	V2
QM9	1	11.068	11.068	0.818	0.818	1.311	1.311
	2.5	1.046	1.046	12.957	12.957	1.042	1.042
	5	1.114	1.114	1.056	1.051	1.242	1.142
	7.5	1.327	1.233	1.157	1.157	1.304	1.304
	10	1.210	1.210	1.072	1.072	1.209	1.109
	25	1.252	1.252	0.858	0.858	1.300	1.300
	50	1.040	1.040	0.994	1.000	1.044	1.044
	75	1.207	1.207	1.263	1.361	1.209	1.209
	100	1.158	1.159	1.167	1.167	1.610	1.610
	125	0.921	0.921	1.073	1.073	1.422	1.423
	1	1.204	1.204	1.330	1.429	1.366	-3.106
	2.5	1.430	1.430	1.300	1.300	1.146	1.146
GDB	5	1.208	1.208	1.168	1.068	1.503	1.503
	7.5	1.169	1.169	1.687	1.687	1.439	1.439
	10	1.259	1.258	1.384	1.384	1.435	1.435
	25	1.506	1.509	1.892	1.892	1.548	1.548
	50	1.159	1.159	1.688	1.688	1.546	1.546
	75	1.258	1.261	1.594	2.195	1.690	1.590
	100	1.463	1.466	1.597	1.595	1.592	1.592
	250	-0.448	0.758	1.502	-0.304	0.198	0.198
	500	0.360	0.360	0.493	2.300	0.405	0.404
	1	1.865	1.865	1.925	1.925	1.935	1.935
	2.5	1.891	1.891	2.533	2.529	2.297	2.297
	5	2.090	2.196	2.461	2.562	2.332	2.332
ZINC	7.5	2.594	2.594	2.802	2.802	2.081	2.180
	10	2.173	2.174	3.121	3.122	2.149	2.147
	25	2.089	2.188	2.618	2.618	2.747	2.746
	50	2.360	2.352	3.279	3.281	2.667	2.660
	75	1.897	1.891	3.329	3.323	2.370	2.370
	100	2.200	2.199	3.302	3.313	2.915	2.915
	250	6.356	6.356	-22.824	-28.264	16.378	21.211
	500	-54.032	-73.345	73.449	42.590	40.645	40.640

The tables display the results for each database subset and every molecular string representation with color codes. The values displayed in ESI-Table 4 are in GB, while the values included in ESI-Table 5 are in MB.

3 Supplementary Figures and Tables

This section contains the comparative results for the three databases, the three molecular string representations, and the encoding methods compared in this paper. The results will be displayed for each model accordingly (ESI-Figure 2 to ESI-Figure 13).

Additionally includes a table for each database and model trained with the eOHE, every table contains information about every subset and molecular string representation used (ESI-Table 6 to ESI-Table 11):

- the length of dictionary ℓ ,
- size of the reduced dictionary p ,
- reduction factor q and
- number of empty tokens m added to dictionary for every subset.

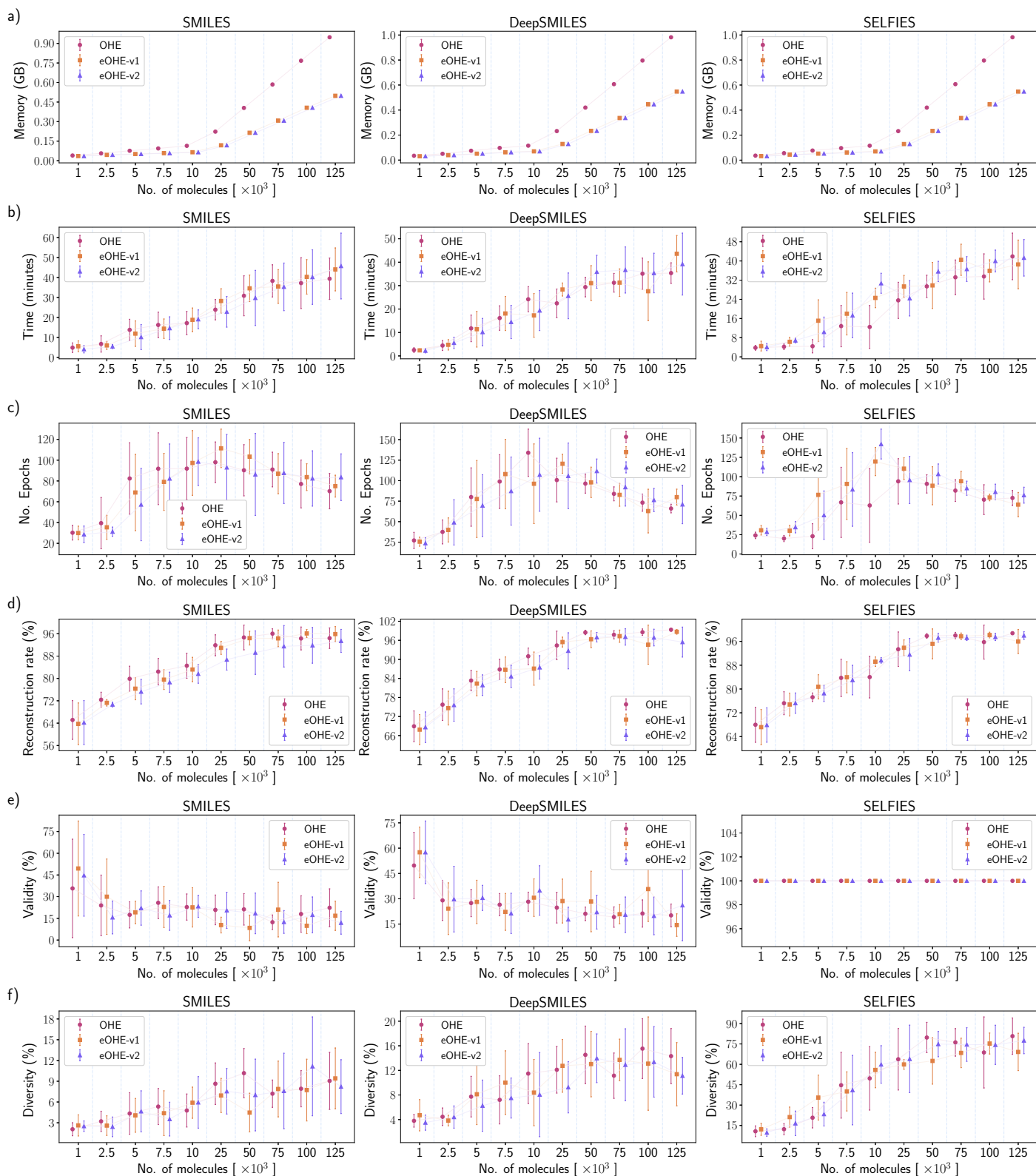


Figure S2 VAE model metric results for QM9 database subsets. In order of appearance per row is (a) Memory usage by the GPU, (b) Training Time, (c) Number of Epochs required to train the model, (d) Reconstruction rate of the model, (e) Percentage of valid molecules sampled from latent space in a sample of 1000 molecules and (f) Diversity of molecules of the same sampled 1000 molecules. The results for SMILES are displayed in all the left subplot, DeepSMILES in all the middle columns subplots and SELFIES for right subplots. The x-axis of every subplot is depicting the amount of molecules used for the training of the model for every subset of the QM9 database. The error bars are displaying in the central value the mean of the respective metric evaluated, while the bars are showing the standard deviations of the respective subsets.

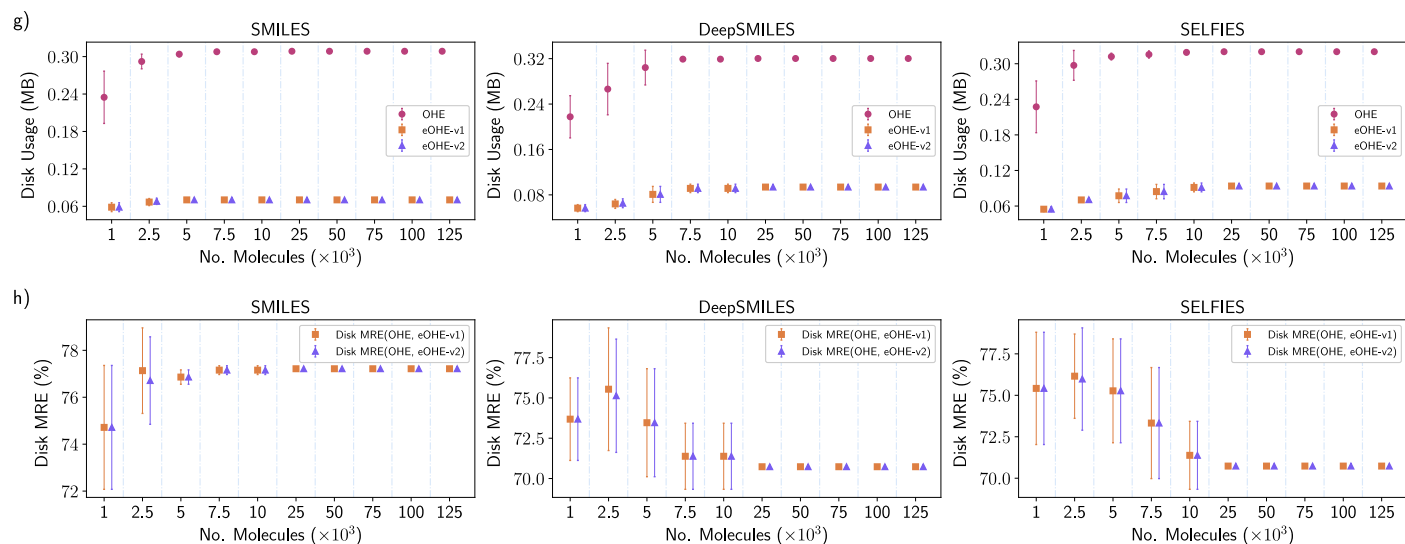


Figure S3 VAE model metric results for QM9 database subsets. In order of appearance per row is (g) Disk space utilization, and (h) disk Memory Reduction Efficiency. The results for SMILES are displayed in all the left subplot, DeepSMILES in all the middle columns subplots and SELFIES for right subplots. The x -axis of every subplot is depicting the amount of molecules used for the training of the model for every subset of the QM9 database. The error bars are displaying in the central value the mean of the respective metric evaluated, while the bars are showing the standard deviations of the respective subsets.

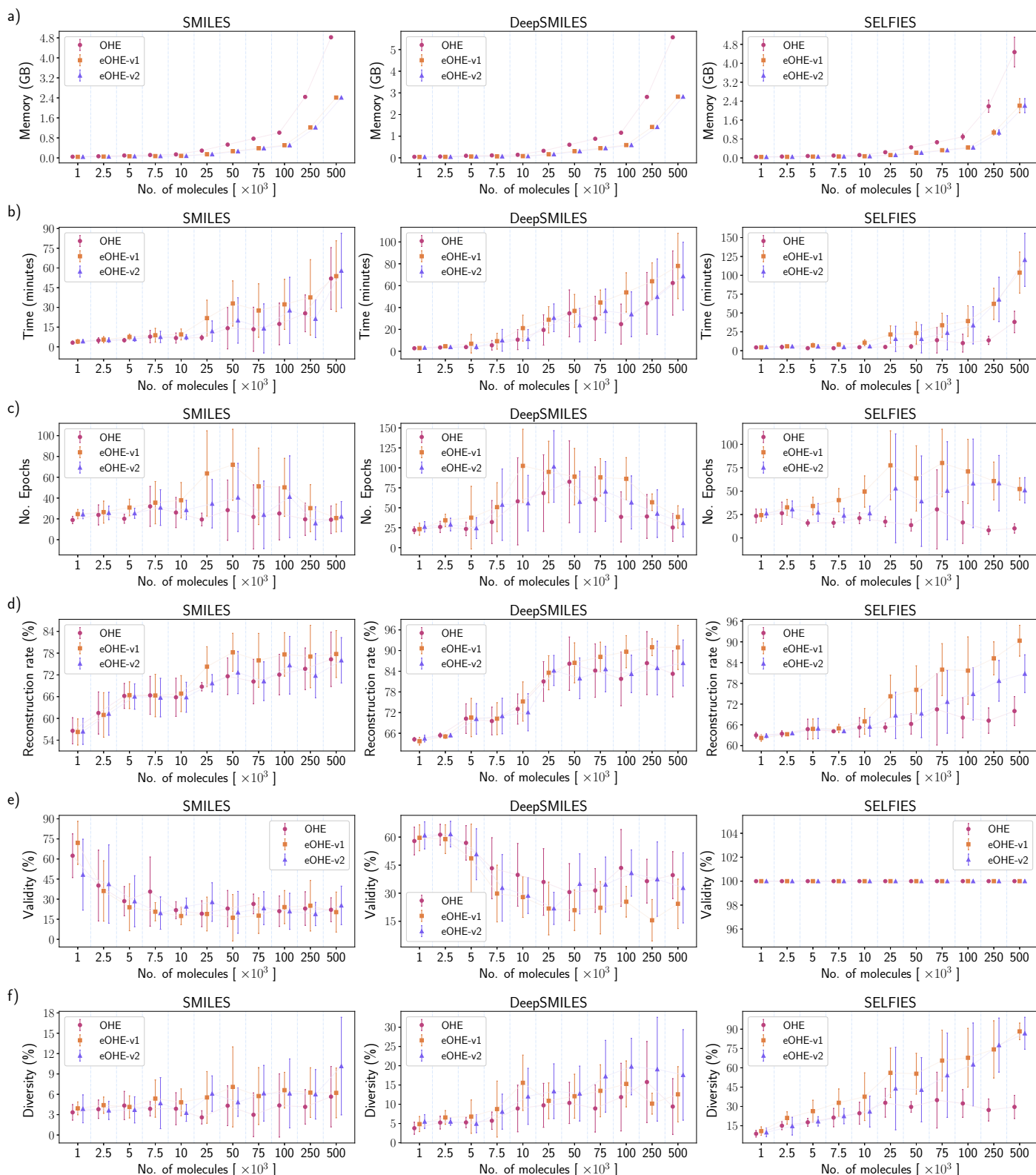


Figure S4 VAE model metric results for GDB-13 database subsets. In order of appearance per row is (a) Memory usage by the GPU, (b) Training Time, (c) Number of Epochs required to train the model, (d) Reconstruction rate of the model, (e) Percentage of valid molecules sampled from latent space in a sample of 1000 molecules and (f) Diversity of molecules of the same sampled 1000 molecules. The results for SMILES are displayed in all the left subplot, DeepSMILES in all the middle columns subplots and SELFIES for right subplots. The x-axis of every subplot is depicting the amount of molecules used for the training of the model for every subset of the GDB-13 database. The error bars are displaying in the central value the mean of the respective metric evaluated, while the bars are showing the standard deviations of the respective subsets.

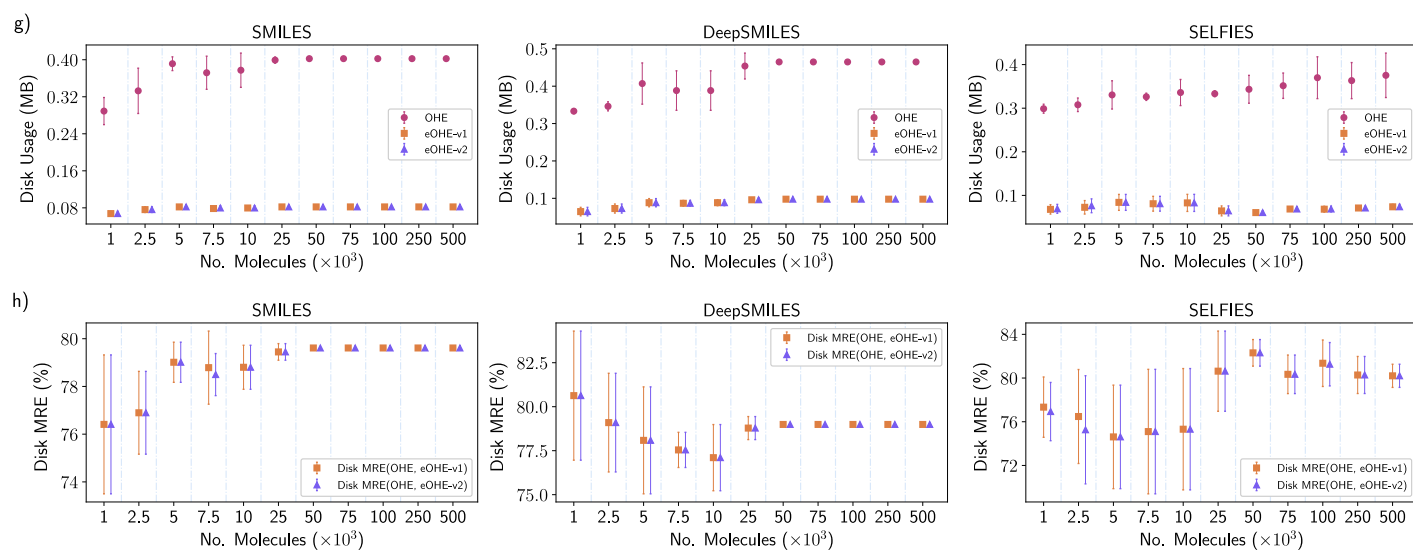


Figure S5 VAE model metric results for GDB-13 database subsets. In order of appearance per row is (g) Disk space utilization, and (h) disk Memory Reduction Efficiency. The results for SMILES are displayed in all the left subplot, DeepSMILES in all the middle columns subplots and SELFIES for right subplots. The x-axis of every subplot is depicting the amount of molecules used for the training of the model for every subset of the GDB database. The error bars are displaying in the central value the mean of the respective metric evaluated, while the bars are showing the standard deviations of the respective subsets.

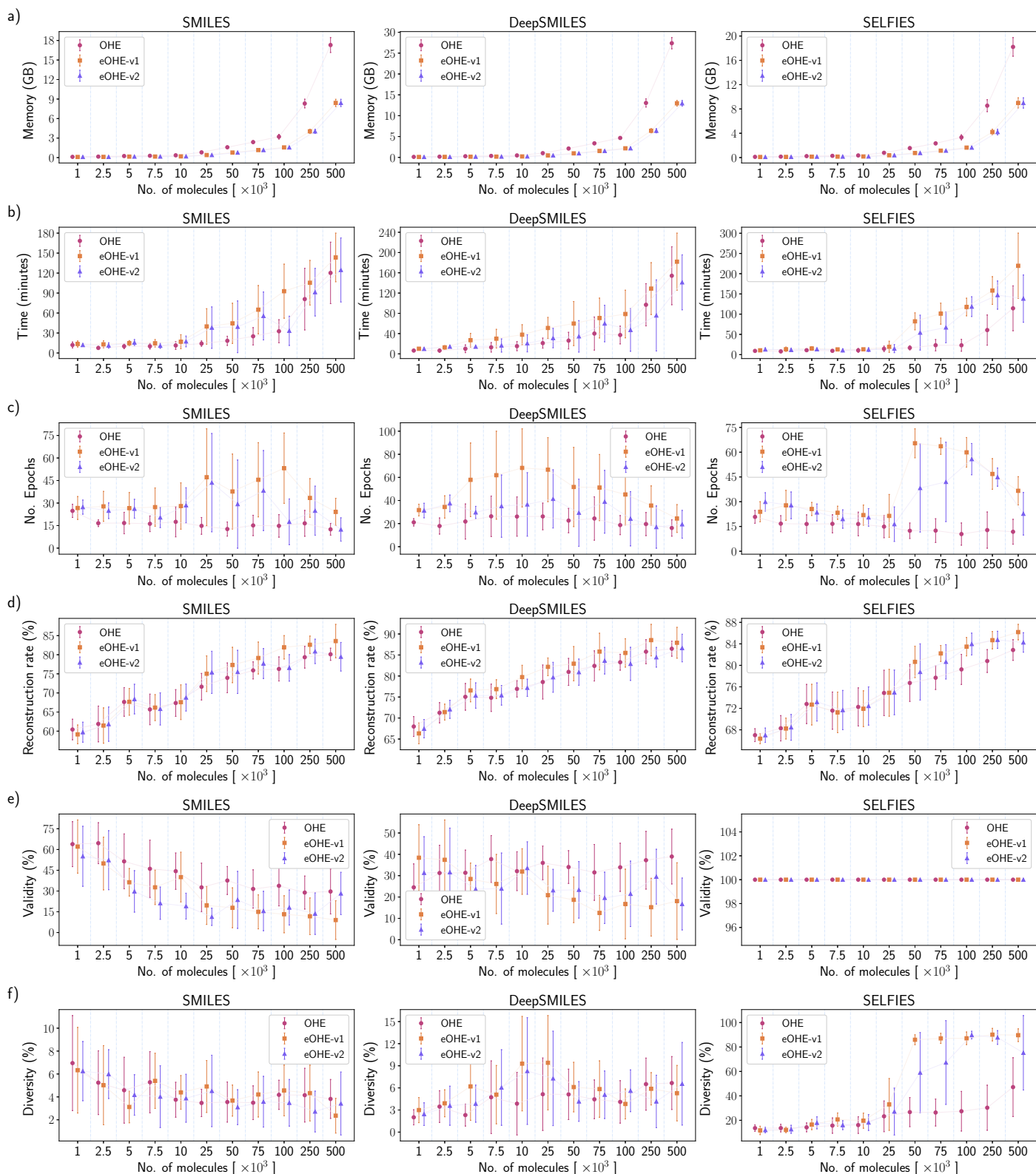


Figure S6 VAE model metric results for ZINC database subsets. In order of appearance per row is (a) Memory usage by the GPU, (b) Training Time, (c) Number of Epochs required to train the model, (d) Reconstruction rate of the model, (e) Percentage of valid molecules sampled from latent space in a sample of 1000 molecules and (f) Diversity of molecules of the same sampled 1000 molecules. The results for SMILES are displayed in all the left subplot, DeepSMILES in all the middle columns subplots and SELFIES for right subplots. The x-axis of every subplot is depicting the amount of molecules used for the training of the model for every subset of the ZINC database. The error bars are displaying in the central value the mean of the respective metric evaluated, while the bars are showing the standard deviations of the respective subsets.

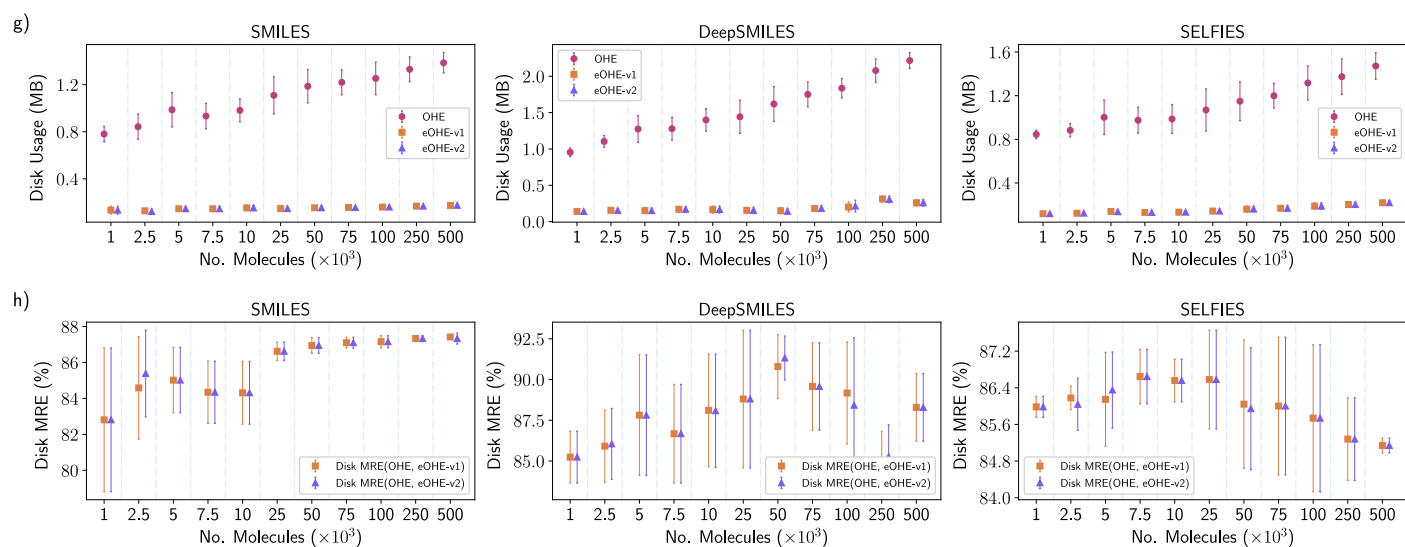


Figure S7 VAE model metric results for ZINC database subsets. In order of appearance per row is (g) Disk space utilization, and (h) disk Memory Reduction Efficiency. The results for SMILES are displayed in all the left subplot, DeepSMILES in all the middle columns subplots and SELFIES for right subplots. The x -axis of every subplot is depicting the amount of molecules used for the training of the model for every subset of the ZINC database. The error bars are displaying in the central value the mean of the respective metric evaluated, while the bars are showing the standard deviations of the respective subsets.

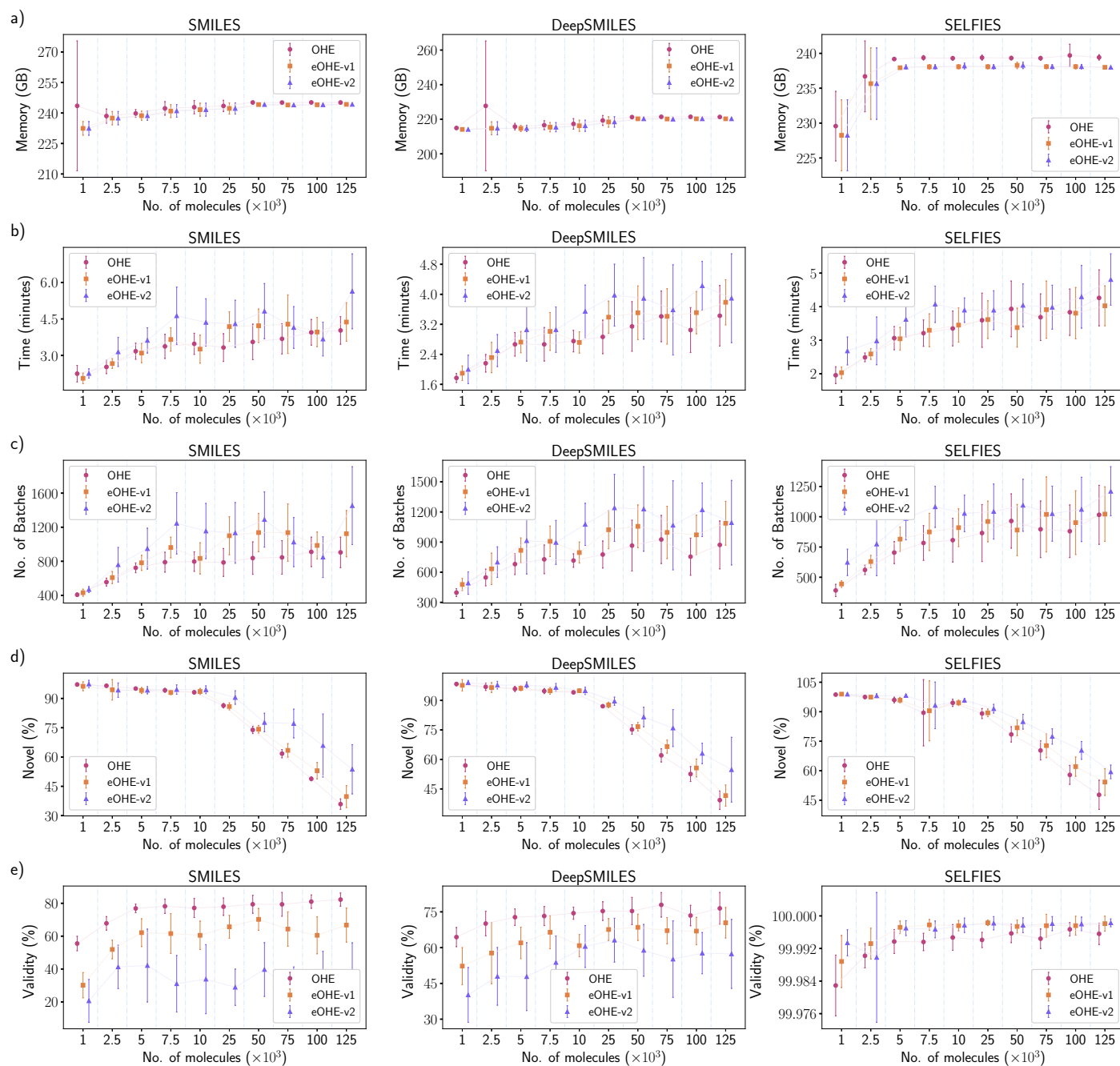


Figure S8 RNN model metric results for QM9 database subsets. In order of appearance per row is (a) Memory usage by the GPU, (b) Training Time, (c) Number of batches required to train the model, (d) percentage of Novel molecules and (e) percentage of Valid molecules sampled from latent space in a sample of 100000 molecules. The results for SMILES are displayed in all the left subplot, DeepSMILES in all the middle columns subplots and SELFIES for right subplots. The x-axis of every subplot is depicting the amount of molecules used for the training of the model for every subset of the QM9 database. The error bars are displaying in the central value the mean of the respective metric evaluated, while the bars are showing the standard deviations of the respective subsets.

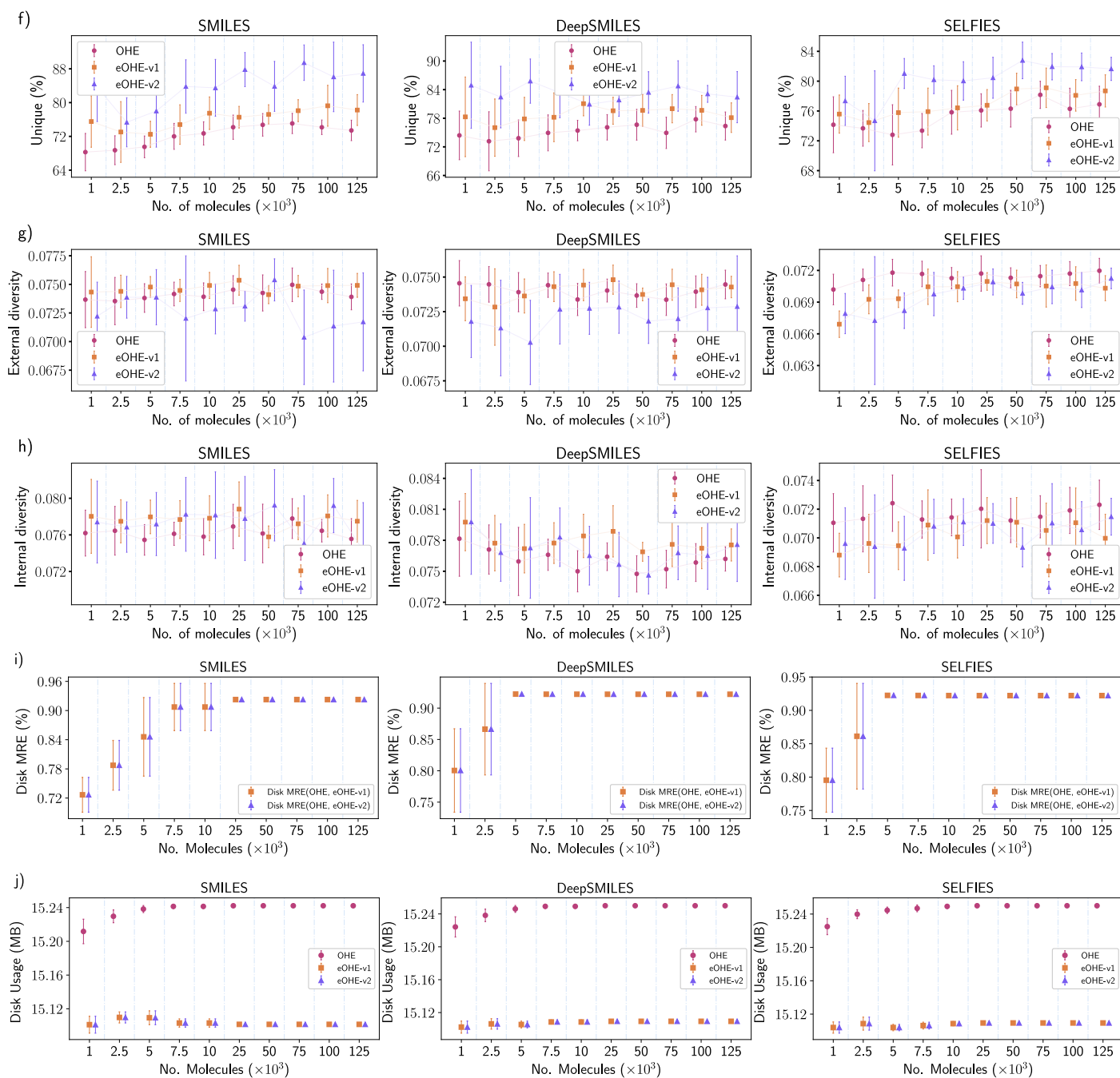


Figure S9 RNN model metric results for QM9 database subsets. In order of appearance per row is (f) percentage of Unique molecules, in a sample of 100000 molecules, (g) External diversity when comparing sampled molecules from training with sampled molecules once the training finish, (h) Internal diversity of sampled molecules, (i) Disk space utilization, and (j) disk Memory Reduction Efficiency. The results for SMILES are displayed in all the left subplot, DeepSMILES in all the middle columns subplots and SELFIES for right subplots. The x-axis of every subplot is depicting the amount of molecules used for the training of the model for every subset of the QM9 database. The error bars are displaying in the central value the mean of the respective metric evaluated, while the bars are showing the standard deviations of the respective subsets.

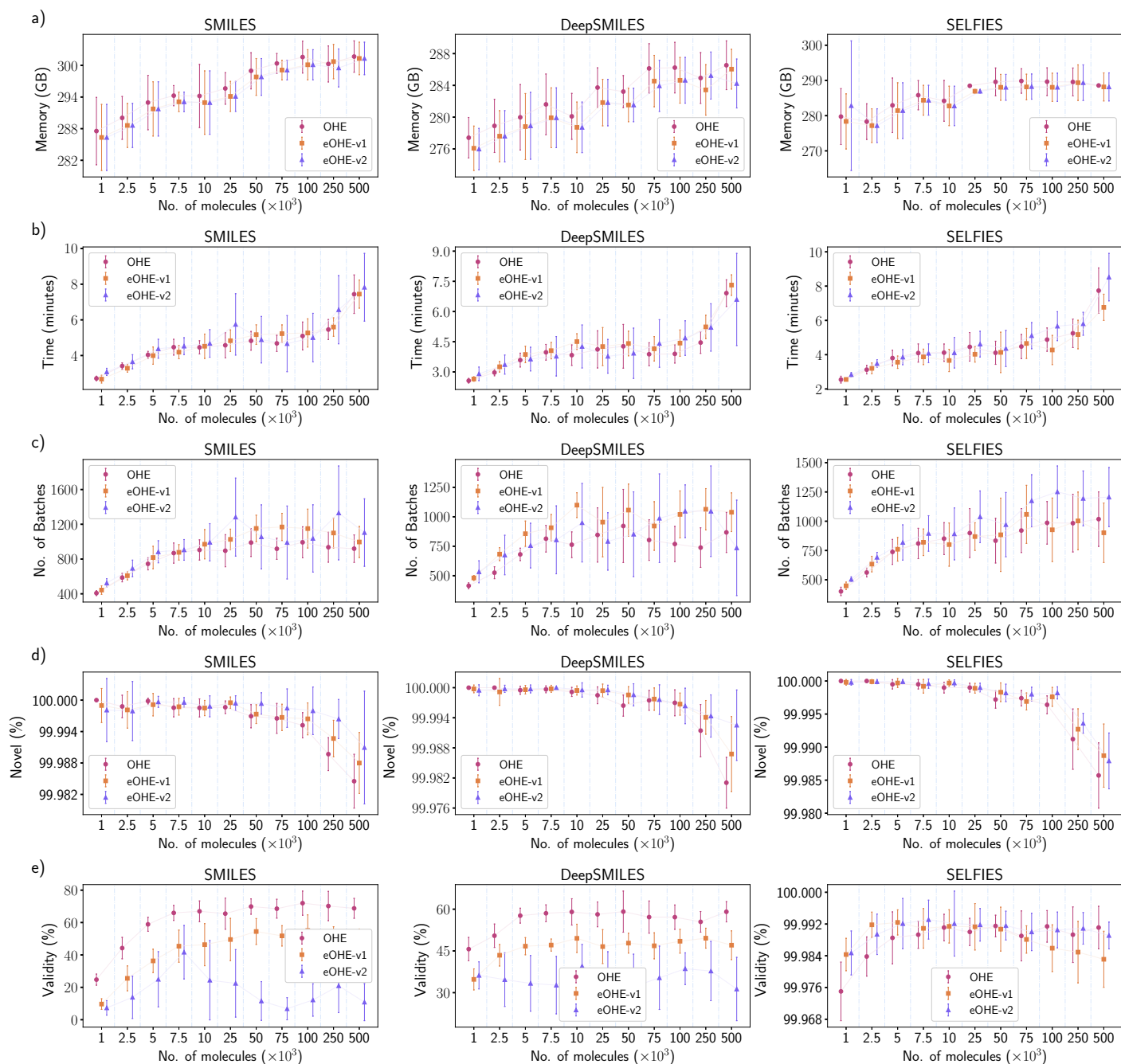


Figure S10 RNN model metric results for GDB-13 database subsets. In order of appearance per row is (a) Memory usage by the GPU, (b) Training Time, (c) Number of batches required to train the model, (d) percentage of Novel molecules and (e) percentage of Valid molecules sampled from latent space in a sample of 100000 molecules. The results for SMILES are displayed in all the left subplot, DeepSMILES in all the middle columns subplots and SELFIES for right subplots. The x-axis of every subplot is depicting the amount of molecules used for the training of the model for every subset of the GDB database. The error bars are displaying in the central value the mean of the respective metric evaluated, while the bars are showing the standard deviations of the respective subsets.

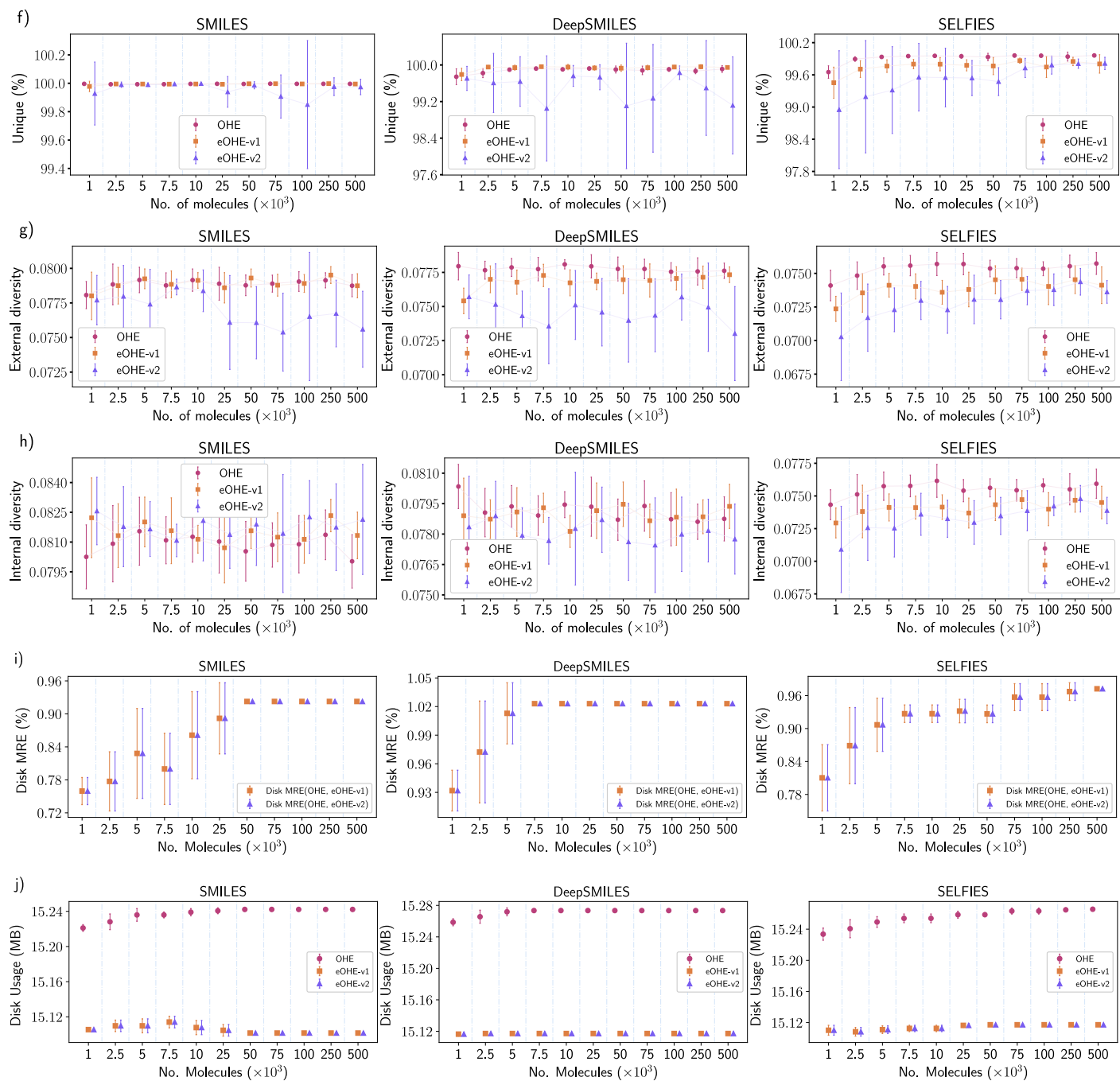


Figure S11 RNN model metric results for GDB-13 database subsets. In order of appearance per row is (f) percentage of Unique molecules, in a sample of 100000 molecules, (g) External diversity when comparing sampled molecules from training with sampled molecules once the training finish, (h) Internal diversity of sampled molecules, (i) Disk space utilization, and (j) disk Memory Reduction Efficiency. The results for SMILES are displayed in all the left subplot, DeepSMILES in all the middle columns subplots and SELFIES for right subplots. The x-axis of every subplot is depicting the amount of molecules used for the training of the model for every subset of the GDB-13 database. The error bars are displaying in the central value the mean of the respective metric evaluated, while the bars are showing the standard deviations of the respective subsets.

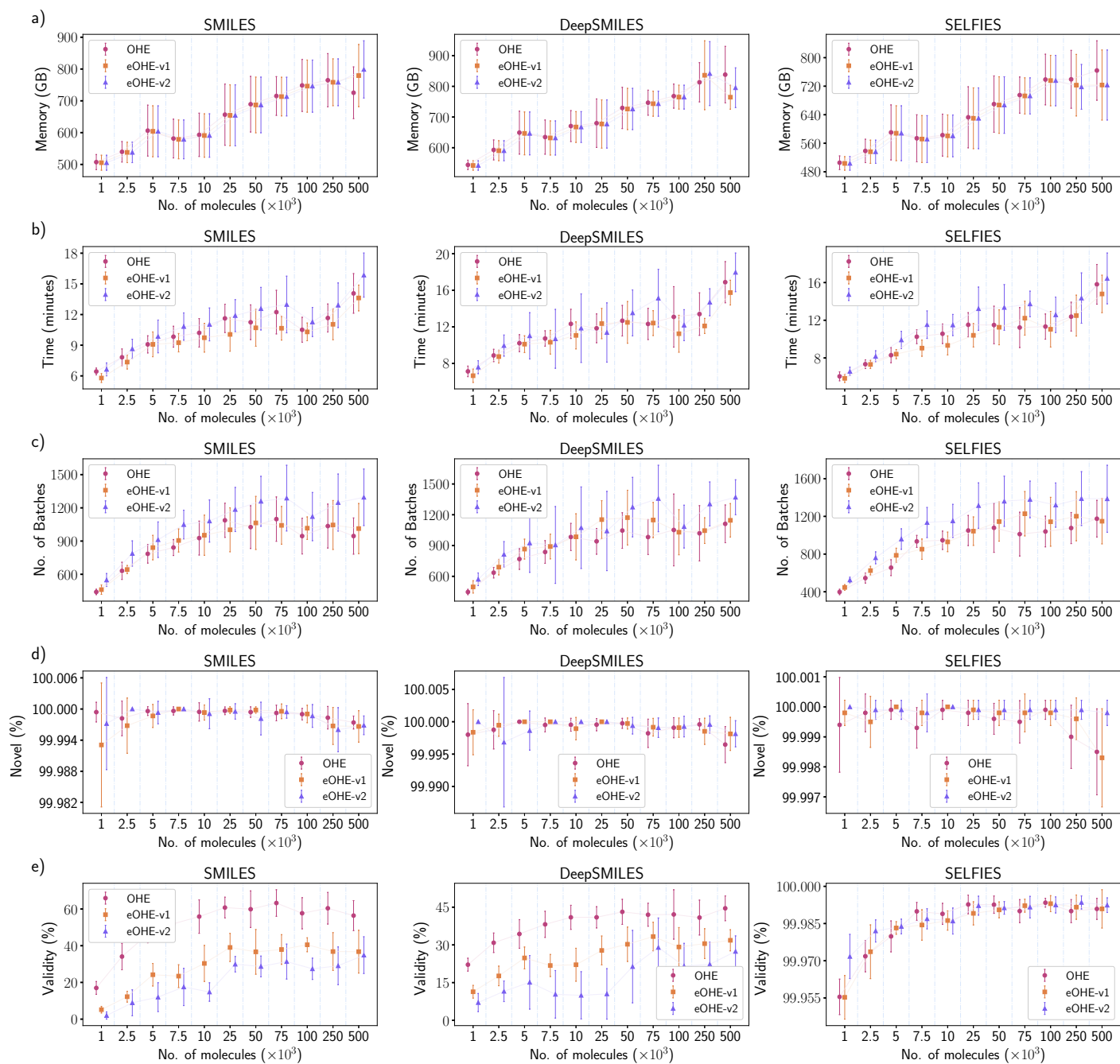


Figure S12 RNN model metric results for ZINC database subsets. In order of appearance per row is (a) Memory usage by the GPU, (b) Training Time, (c) Number of batches required to train the model, (d) percentage of Novel molecules and (e) percentage of Valid molecules sampled from latent space in a sample of 100000 molecules. The results for SMILES are displayed in all the left subplot, DeepSMILES in all the middle columns subplots and SELFIES for right subplots. The x-axis of every subplot is depicting the amount of molecules used for the training of the model for every subset of the ZINC database. The error bars are displaying in the central value the mean of the respective metric evaluated, while the bars are showing the standard deviations of the respective subsets.

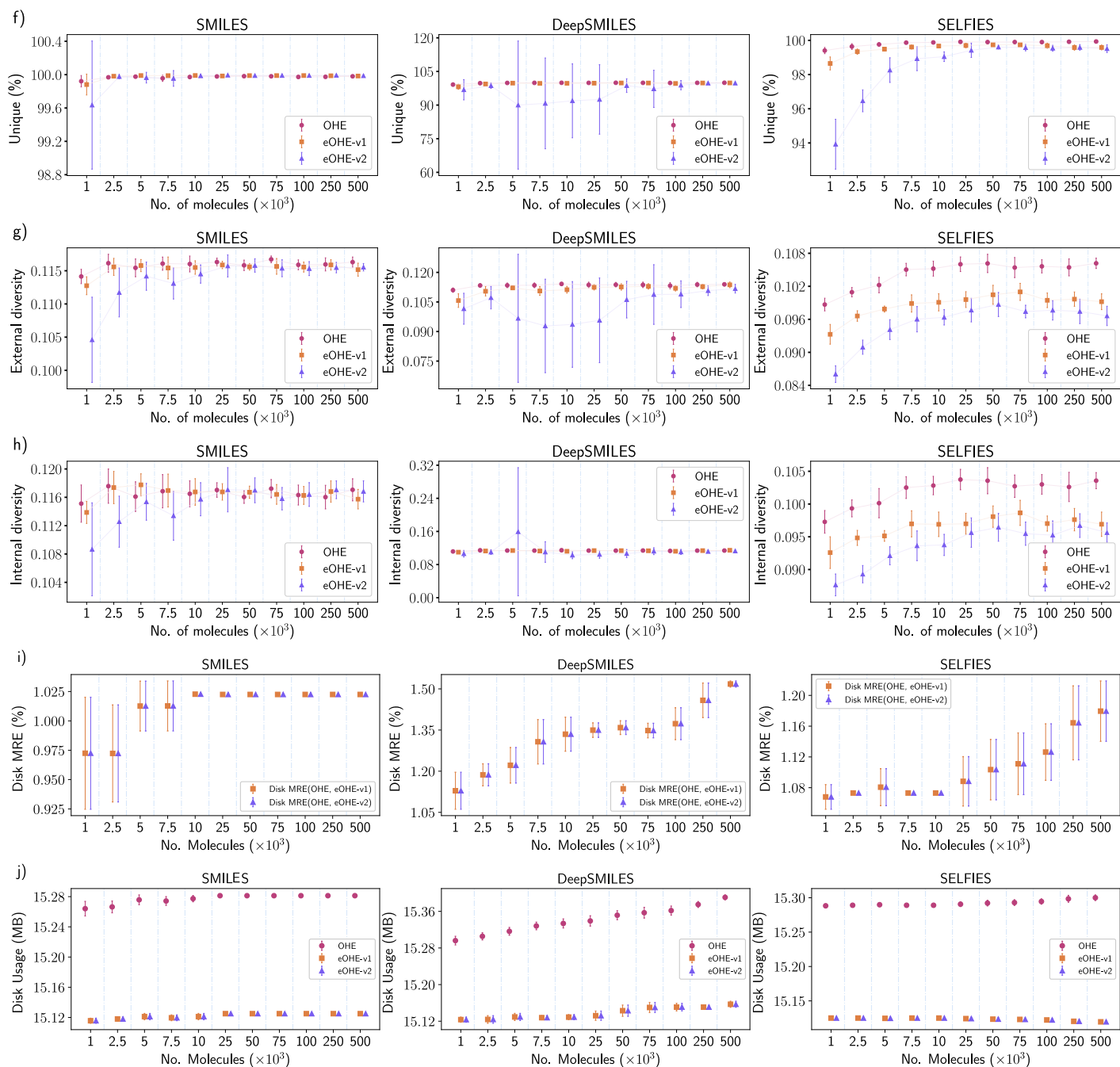


Figure S13 RNN model metric results for ZINC database subsets. In order of appearance per row is (f) percentage of Unique molecules, in a sample of 100000 molecules, (g) External diversity when comparing sampled molecules from training with sampled molecules once the training finish, (h) Internal diversity of sampled molecules, (i) Disk space utilization, and (j) disk Memory Reduction Efficiency. The results for SMILES are displayed in all the left subplot, DeepSMILES in all the middle columns subplots and SELFIES for right subplots. The x-axis of every subplot is depicting the amount of molecules used for the training of the model for every subset of the ZINC database. The error bars are displaying in the central value the mean of the respective metric evaluated, while the bars are showing the standard deviations of the respective subsets.

Table S10 Explicit values for length of dictionary ℓ , size of the reduced dictionary p , reduction factor q and number of empty tokens m added to dictionary, displayed for eOHE-v1 and eOHE-v2, for every subset of ZINC database using VAE model. The values are displayed as “(ℓ , p , q , m)” for every molecular string representation using a color code: ■ SMILES, ■ DeepSMILES, ■ SELFIES.

	No. of Molecules ($\times 10^3$)															
Idx	1	2.5	5	7.5	10	25	50	75	100	250	500					
0	(29, 3, 10, 1)	(29, 3, 10, 1)	(31, 4, 8, 1)	(29, 3, 10, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)					
1	(30, 3, 10, 0)	(28, 4, 7, 0)	(31, 4, 8, 1)	(30, 3, 10, 0)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)					
2	(28, 4, 7, 0)	(30, 3, 10, 0)	(30, 3, 10, 0)	(30, 3, 10, 0)	(30, 3, 10, 0)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)					
3	(30, 3, 10, 0)	(28, 4, 7, 0)	(29, 3, 10, 1)	(30, 3, 10, 0)	(30, 3, 10, 0)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)					
4	(27, 3, 9, 0)	(29, 3, 10, 1)	(29, 3, 10, 1)	(31, 4, 8, 1)	(30, 3, 10, 0)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)					
5	(28, 4, 7, 0)	(29, 3, 10, 1)	(30, 3, 10, 0)	(31, 4, 8, 1)	(30, 3, 10, 0)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)					
6	(30, 3, 10, 0)	(31, 4, 8, 1)	(30, 3, 10, 0)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)					
7	(27, 3, 9, 0)	(28, 4, 7, 0)	(31, 4, 8, 1)	(30, 3, 10, 0)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)					
8	(30, 3, 10, 0)	(30, 3, 10, 0)	(31, 4, 8, 1)	(29, 3, 10, 1)	(30, 3, 10, 0)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)					
9	(29, 3, 10, 1)	(29, 3, 10, 1)	(31, 4, 8, 1)	(30, 3, 10, 0)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)					
0	(32, 4, 8, 0)	(34, 2, 17, 0)	(36, 4, 9, 0)	(37, 2, 19, 1)	(37, 2, 19, 1)	(39, 3, 13, 0)	(40, 4, 10, 0)	(41, 6, 7, 1)	(43, 4, 11, 1)	(44, 4, 11, 0)	(45, 5, 9, 0)					
1	(35, 5, 7, 0)	(33, 3, 11, 0)	(36, 4, 9, 0)	(37, 2, 19, 1)	(38, 2, 19, 0)	(37, 2, 19, 1)	(38, 2, 19, 0)	(41, 6, 7, 1)	(42, 6, 7, 0)	(42, 6, 7, 0)	(45, 5, 9, 0)					
2	(33, 3, 11, 0)	(33, 3, 11, 0)	(35, 5, 7, 0)	(38, 2, 19, 0)	(39, 3, 13, 0)	(40, 4, 10, 0)	(42, 6, 7, 0)	(41, 6, 7, 1)	(41, 6, 7, 1)	(44, 4, 11, 0)	(44, 4, 11, 0)					
3	(34, 2, 17, 0)	(34, 2, 17, 0)	(34, 2, 17, 0)	(38, 2, 19, 0)	(38, 2, 19, 0)	(38, 2, 19, 0)	(39, 3, 13, 0)	(41, 6, 7, 1)	(41, 6, 7, 1)	(42, 6, 7, 0)	(46, 6, 8, 2)					
4	(31, 4, 8, 1)	(35, 5, 7, 0)	(36, 4, 9, 0)	(36, 4, 9, 0)	(36, 4, 9, 0)	(39, 3, 13, 0)	(41, 6, 7, 1)	(40, 4, 10, 0)	(42, 6, 7, 0)	(44, 4, 11, 0)	(44, 4, 11, 0)					
5	(33, 3, 11, 0)	(34, 2, 17, 0)	(37, 2, 19, 1)	(38, 2, 19, 0)	(38, 2, 19, 0)	(37, 2, 19, 1)	(39, 3, 13, 0)	(40, 4, 10, 0)	(40, 4, 10, 0)	(42, 6, 7, 0)	(45, 5, 9, 0)					
6	(32, 4, 8, 0)	(36, 4, 9, 0)	(37, 2, 19, 1)	(37, 2, 19, 1)	(37, 2, 19, 1)	(37, 2, 19, 1)	(40, 4, 10, 0)	(42, 6, 7, 0)	(41, 6, 7, 1)	(43, 4, 11, 1)	(45, 5, 9, 0)					
7	(32, 4, 8, 0)	(33, 3, 11, 0)	(34, 2, 17, 0)	(38, 2, 19, 0)	(38, 2, 19, 0)	(37, 2, 19, 1)	(39, 3, 13, 0)	(42, 6, 7, 0)	(41, 6, 7, 1)	(43, 4, 11, 1)	(46, 6, 8, 2)					

Table S11 Explicit values for length of dictionary ℓ , size of the reduced dictionary p , reduction factor q and number of empty tokens m added to dictionary, displayed for eOHE-v1 and eOHE-v2, for every subset of ZINC database using RNN model. The values are displayed as “(ℓ , p , q , m)” for every molecular string representation using a color code: ■ SMILES, ■ DeepSMILES, ■ SELFIES.

	No. of Molecules ($\times 10^3$)										
Idx	1	2.5	5	7.5	10	25	50	75	100	250	500
0	(29, 3, 10, 1)	(29, 3, 10, 1)	(31, 4, 8, 1)	(29, 3, 10, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)
1	(30, 3, 10, 0)	(28, 4, 7, 0)	(31, 4, 8, 1)	(30, 3, 10, 0)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)
2	(28, 4, 7, 0)	(30, 3, 10, 0)	(30, 3, 10, 0)	(30, 3, 10, 0)	(30, 3, 10, 0)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)
3	(30, 3, 10, 0)	(28, 4, 7, 0)	(29, 3, 10, 1)	(30, 3, 10, 0)	(30, 3, 10, 0)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)
4	(27, 3, 9, 0)	(29, 3, 10, 1)	(29, 3, 10, 1)	(31, 4, 8, 1)	(30, 3, 10, 0)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)
5	(28, 4, 7, 0)	(29, 3, 10, 1)	(30, 3, 10, 0)	(31, 4, 8, 1)	(30, 3, 10, 0)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)
6	(30, 3, 10, 0)	(31, 4, 8, 1)	(30, 3, 10, 0)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)
7	(27, 3, 9, 0)	(28, 4, 7, 0)	(31, 4, 8, 1)	(30, 3, 10, 0)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)
8	(30, 3, 10, 0)	(30, 3, 10, 0)	(31, 4, 8, 1)	(29, 3, 10, 1)	(30, 3, 10, 0)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)
9	(29, 3, 10, 1)	(29, 3, 10, 1)	(31, 4, 8, 1)	(30, 3, 10, 0)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)	(31, 4, 8, 1)
0	(32, 4, 8, 0)	(34, 2, 17, 0)	(36, 4, 9, 0)	(37, 2, 19, 1)	(37, 2, 19, 1)	(39, 3, 13, 0)	(40, 4, 10, 0)	(41, 6, 7, 1)	(43, 4, 11, 1)	(44, 4, 11, 0)	(45, 5, 9, 0)
1	(35, 5, 7, 0)	(33, 3, 11, 0)	(36, 4, 9, 0)	(37, 2, 19, 1)	(38, 2, 19, 0)	(37, 2, 19, 1)	(38, 2, 19, 0)	(41, 6, 7, 1)	(42, 6, 7, 0)	(42, 6, 7, 0)	(45, 5, 9, 0)
2	(33, 3, 11, 0)	(33, 3, 11, 0)	(35, 5, 7, 0)	(38, 2, 19, 0)	(39, 3, 13, 0)	(40, 4, 10, 0)	(42, 6, 7, 0)	(41, 6, 7, 1)	(41, 6, 7, 1)	(44, 4, 11, 0)	(44, 4, 11, 0)
3	(34, 2, 17, 0)	(34, 2, 17, 0)	(34, 2, 17, 0)	(38, 2, 19, 0)	(38, 2, 19, 0)	(38, 2, 19, 0)	(39, 3, 13, 0)	(41, 6, 7, 1)	(41, 6, 7, 1)	(42, 6, 7, 0)	(46, 6, 8, 2)
4	(31, 4, 8, 1)	(35, 5, 7, 0)	(36, 4, 9, 0)	(36, 4, 9, 0)	(36, 4, 9, 0)	(39, 3, 13, 0)	(41, 6, 7, 1)	(40, 4, 10, 0)	(42, 6, 7, 0)	(44, 4, 11, 0)	(44, 4, 11, 0)
5	(33, 3, 11, 0)	(34, 2, 17, 0)	(37, 2, 19, 1)	(38, 2, 19, 0)	(38, 2, 19, 0)	(37, 2, 19, 1)	(39, 3, 13, 0)	(40, 4, 10, 0)	(40, 4, 10, 0)	(42, 6, 7, 0)	(45, 5, 9, 0)
6	(32, 4, 8, 0)	(36, 4, 9, 0)	(37, 2, 19, 1)	(37, 2, 19, 1)	(37, 2, 19, 1)	(37, 2, 19, 1)	(40, 4, 10, 0)	(42, 6, 7, 0)	(41, 6, 7, 1)	(43, 4, 11, 1)	(45, 5, 9, 0)
7	(32, 4, 8, 0)	(33, 3, 11, 0)	(34, 2, 17, 0)	(38, 2, 19, 0)	(38, 2, 19, 0)	(37, 2, 19, 1)	(39, 3, 13, 0)	(42, 6, 7, 0)	(41, 6, 7, 1)	(43, 4, 11, 1)	(46, 6, 8, 2)
8	(33, 3, 11, 0)	(35, 5, 7, 0)	(35, 5, 7, 0)	(35, 5, 7, 0)	(36, 4, 9, 0)	(39, 3, 13, 0)	(41, 6, 7, 1)	(42, 6, 7, 0)	(43, 4, 11, 1)	(43, 4, 11, 1)	(45, 5, 9, 0)
9	(34, 2, 17, 0)	(34, 2, 17, 0)	(35, 5, 7, 0)	(36, 4, 9, 0)	(40, 4, 10, 0)	(41, 6, 7, 1)	(41, 6, 7, 1)	(37, 2, 19, 1)	(39, 3, 13, 0)	(43, 4, 11, 1)	(45, 5, 9, 0)
0	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(33, 3, 11, 0)	(34, 2, 17, 0)	(33, 3, 11, 0)
1	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(33, 3, 11, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(33, 3, 11, 0)	(33, 3, 11, 0)
2	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(33, 3, 11, 0)	(34, 2, 17, 0)
3	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(33, 3, 11, 0)	(32, 4, 8, 0)	(34, 2, 17, 0)
4	(31, 4, 8, 1)	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(33, 3, 11, 0)	(33, 3, 11, 0)	(33, 3, 11, 0)	(33, 3, 11, 0)	(33, 3, 11, 0)
5	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(33, 3, 11, 0)	(32, 4, 8, 0)	(33, 3, 11, 0)	(33, 3, 11, 0)	(34, 2, 17, 0)	(33, 3, 11, 0)
6	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(33, 3, 11, 0)	(33, 3, 11, 0)	(33, 3, 11, 0)	(33, 3, 11, 0)
7	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(33, 3, 11, 0)	(33, 3, 11, 0)	(32, 4, 8, 0)	(34, 2, 17, 0)	(34, 2, 17, 0)
8	(32, 4, 8, 0)	(32, 4, 8, 0)	(33, 3, 11, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(33, 3, 11, 0)	(32, 4, 8, 0)	(33, 3, 11, 0)	(33, 3, 11, 0)	(33, 3, 11, 0)
9	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(32, 4, 8, 0)	(33, 3, 11, 0)	(33, 3, 11, 0)	(33, 3, 11, 0)	(33, 3, 11, 0)	(34, 2, 17, 0)

Notes and references

- 1 M. Krenn, F. Häse, A. Nigam, P. Friederich and A. Aspuru-Guzik, *Mach. Learn.: Sci. Technol.*, 2020, **1**, 045024.
- 2 M. A. Skinnider, R. G. Stacey, D. S. Wishart and L. J. Foster, *Nat. Mach. Intell.*, 2021, **3**, 759–770.
- 3 H. F. Soon, A. Amir and S. N. Azemi, *J. Phys.: Conf. Ser.*, 2021, **1755**, 012030.
- 4 G. Aguiar, B. Krawczyk and A. Cano, *Mach. Learn.*, 2023.