Electronic Supplementary Material (ESI) for Molecular Systems Design & Engineering.

Supplementary Materials for:

Featurization Strategies for Polymer Sequence or Composition Design by Machine Learning

Roshan A. Patel, Carlos H. Borca, Michael A. Webb*

* Corresponding author. Email: mawebb@princeton.edu

The PDF file includes:

Sections S1 to S3

Other Supplementary Material for this manuscript includes the following:

DatasetA_metadata.csv

DatasetB_metadata.csv

DatasetC_metadata.csv

DatasetD_metadata.csv

Dataset_A_sequences.txt

Labels.csv

Model_Performances.csv

S1 Simulation Details and Calculated Properties

This section provides information about the simulations performed of IDP sequences in Dataset A and how properties were computed.

In-house molecular dynamics simulations were performed with the HPS model using the LAMMPS simulation package. Simulations were preceded by an energy minimization and run for 10^9 fs using timesteps of 10 fs, thermostatted at 300 K using the Langevin thermostat with a damping constant of 1 ps. Thermodynamic quantities, used for the calculation of heat capacity C_v, were obtained in intervals of 100 ps. Atom coordinates, used for the calculation radius of gyration R_g and decorrelation time τ_N, were obtained in intervals of 5 ps. The following equations were used to calculate R_g, C_v, and τ_N:

$$R_\mathrm{g} := \langle R_\mathrm{g}^2 \rangle^{1/2} = \left(\frac{1}{N} \sum_{i=1}^{N} \langle (R_i - R_{CM})^2 \rangle \right)^{\frac{1}{2}}$$

where N is the total number of atoms, and R_i and R_{CM} are the position of atom i and center of mass of all atoms in the system respectively,

$$C_v := \langle C_v \rangle = \frac{\langle E^2 \rangle - \langle E \rangle^2}{k_b T^2}$$

where E is the total internal energy of the system,

$$\tau_N := \langle \tau_N \rangle = \int_0^\infty \langle \delta R(t) \delta R(0) \rangle \, dt \, , \delta R(t) = R_{i=N}(t) - R_{i=1}(t)$$

where $R_{i=N}(t)$ and $R_{i=1}(t)$ are the end positions of the polymer at a given time t. The integral was approximated by fitting the end-to-end time autocorrelation function to a Kohlrausch–Williams–Watts (KWW) function and performing an analytical integration.

S2 Model Architectures and Hyperparameters

This section broadly covers hyperparameters and their associated considerations in evaluating featurization strategies. Sections S2.1-S2.4 provide details on the different neural network architectures employed and their associated hyperparameters. The indication of an “optional” layer means that the presence of the layer itself, and all associated parameters, is a hyperparameter. The indication of an “optional, conditional” layer means that its presence is again a hyperparameter but is conditional on the presence of another indicated hyperparameter. Section S2.5 probes the sensitivity of model performance trained to architecture and training hyperparameters for fixed featurization strategies.

S2.1 Densely Connected Neural Network

L1: Dense Layer

- Size: [10 – 750] intervals of 20
- Dropout: [0.0 – 0.8] intervals of 0.1
- Activation: ReLU

L2: Dense Layer (Optional)

- Size: [10 – 750] intervals of 20
- Dropout: [0.0 – 0.8] intervals of 0.1
- Activation: ReLU

L3: Dense Layer (Optional, Conditional on L2)

- Size: [10 – 750] intervals of 20
- Dropout: [0.0 – 0.8] intervals of 0.1
- Activation: ReLU

S2.2 1-D Convolutional Neural Network

Conv1: 1-D Convolutional Layer

- Filter: [8-64] intervals of 8
- Kernel Width: [5,25] intervals of 5

P1: 1-D Pooling Layer (Optional)

Type: Max, Average

Size: [3-9] intervals of 2

Conv2: 1-D Convolutional Layer (Optional)

Filter: [8-64] intervals of 8

Kernel Width: [5,25] intervals of 5

P2: 1-D Pooling Layer (Optional)

Type: Max, Average

Size: [3-9] intervals of 2

Flatten: Flattening operation

L1: Dense Layer

Size: [10 – 750] intervals of 20

Dropout: [0.0 – 0.8] intervals of 0.1

Activation: ReLU

L2: Dense Layer (Optional)

Size: [10 – 750] intervals of 20

Dropout: [0.0 – 0.8] intervals of 0.1

Activation: ReLU

S2.3 Graph Convolutional Neural Networks

GL1: Graph Convolutional Layer

Type: GCN, GAT

Size: [2, 42] intervals of 4

GL2: Graph Convolutional Layer (Optional)

Type: GCN, GAT

Size: [2, 42] intervals of 4

P1: Pooling

Type: Sum, Average

L1: Dense Layer

Size: [10 – 750] intervals of 20

Dropout: [0.0 – 0.8] intervals of 0.1

Activation: ReLU

L2: Dense Layer (Optional)

Size: [10 – 750] intervals of 20

Dropout: [0.0 – 0.8] intervals of 0.1

Activation: ReLU

S2.4 Long short-term memory cells

B- LSTM1: Bidirectional Long Short-Term Memory Cell

Size: [5-20] intervals of 5

LSTM2: Long Short-Term Memory Cell (Optional)

Size: [5-20] intervals of 5

Flat: Flattening operation

L1: Dense Layer

Size: [10 – 750] intervals of 20

Dropout: [0.0 – 0.8] intervals of 0.1

Activation: ReLU

L2: Dense Layer (Optional)

Size: [10 – 750] intervals of 20

Dropout: [0.0 – 0.8] intervals of 0.1

Activation: ReLU

Common Hyperparameters and Training Settings

Batch Size = [32,64,128,256] *** This interval number was scaled by 10 for models trained on dataset D due to being an order of magnitude larger than the other datasets.

Learning Rates = [0.001, 0.005,0.01]

Optimizer = Adam

Epochs = 400

Early Stopping Employed, 50 epochs of patience

Validation Split = 15% Training data

During training, it was found that transforming inputs and outputs occasionally helped improve performances of models. The nature of the transformation or whether it was performed is detailed in the Model Performances csv provided in the supporting information.

S2.5 Hyperparameter Sensitivity Analysis

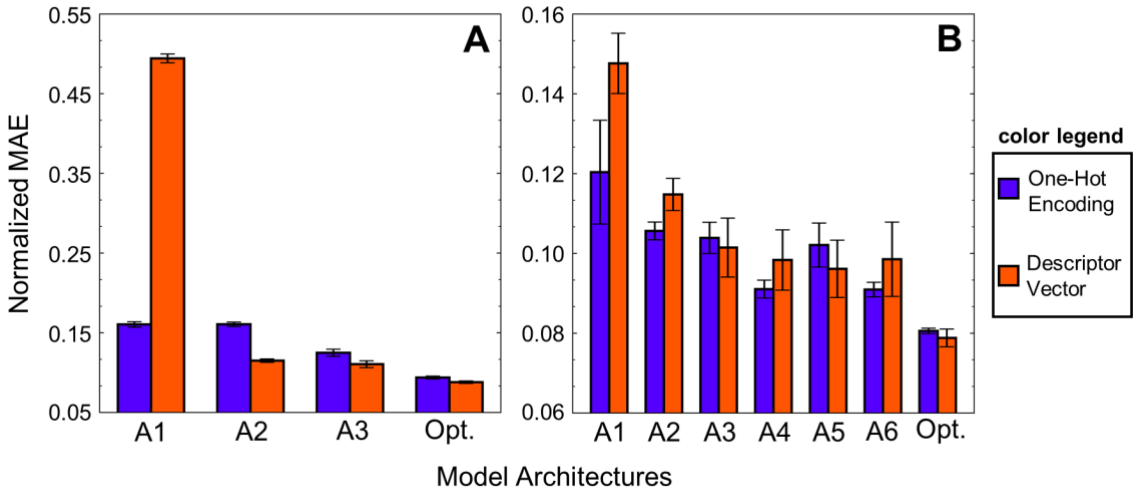

Fig S1: Comparison of CU fingerprints made with different, fixed model architectures for the prediction task in Dataset B. The panels display model performances constructed from **(A)** scaled fingerprint and **(B)** sequence tensor featurization paradigms. Both bars above an entry labeled by A* are performances obtained from the same architecture. Each bar above an entry labeled by Opt. are the performances of hyperparameter-optimized models for each CU vector representation. Hyperparameters associated with specific entries are listed in Table S1.

Model architectures and other training hyperparameters were optimized to construct fair comparisons between the utility of the featurization strategies explored in this study. Ultimately, the selection of a particular featurization strategy can be viewed as its own hyperparameter during model development. In this setting, hyperparameters would ideally be simultaneously co-optimized to construct the optimal model, defined by a fixed featurization strategy, model architecture, and learning procedure. From here, one can reasonably conclude that the selected featurization strategy is the best possible representation of the current data.

Fig. S1 shows that both the absolute and relative performance of models can vary significantly as a function of model architecture. For example, entry A1 of panel A suggests using a descriptor vector for CU representation in the scaled fingerprint paradigm is significantly worse than using a OHE vector. Entries A2, A3, and the hyperparameter-optimized models suggest the opposite. Similarly in panel B, entries A1 and A2 suggest the descriptor vector is worse than OHE for CNN model performance trained on a sequence tensor representation, whereas the remaining entries suggest there is no significant difference. To control for this sensitivity, we only compare hyperparameter-optimized models in the main text. For a given architecture, the domain of hyperparameter optimization is kept fixed when training models to enable facile

comparisons of different CU fingerprints. Therefore, the number of hyperparameters associated with an architecture (DNN, CNN, LSTM, GCN) is considered as part of the overall featurization strategy.

Architecture	Fold	Panel	1-D Conv	Pool	1-D Conv-2	Pool2	Flatten	DNN-1	Dropout-1	DNN-2	Dropout-2	DNN-3	Dropout-3	BS	LR
A1	1-5	A	-	-	-	-	-	20, relu	0.1	20, relu	0.1	-	-	32	0.001
A2	1-5	A	-	-	-	-	-	250, relu	0.1	250, relu	0.1	-	-	32	0.001
A3	1-5	A	-	-	-	-	-	500, relu	0.1	500, relu	0.1	-	-	32	0.001
Opt, OHE	1	A	-	-	-	-	-	740, relu	0.5	700, relu	0.2	420, relu	0	16	0.005
Opt, OHE	2	A	-	-	-	-	-	600, relu	0.2	560, relu	0.5	140, relu	0.1	16	0.001
Opt, OHE	3	A	-	-	-	-	-	460, relu	0.1	360, relu	0.5	380, relu	0.5	64	0.01
Opt, OHE	4	A	-	-	-	-	-	740, relu	0.5	400, relu	0.3	660, relu	0.5	32	0.005
Opt, OHE	5	A	-	-	-	-	-	700, relu	0.3	300, relu	0.1	740, relu	0.7	32	0.005
Opt, Desc. Vec.	1	A	-	-	-	-	-	540, relu	0.4	500, relu	0	320, relu	0.1	16	0.001
Opt, Desc. Vec.	2	A	-	-	-	-	-	440, relu	0	740, relu	0.4	240, relu	0.1	16	0.005
Opt, Desc. Vec.	3	A	-	-	-	-	-	540, relu	0.2	680, relu	0	520, relu	0.1	16	0.001
Opt, Desc. Vec.	4	A	-	-	-	-	-	740, relu	0	740, relu	0.2	680, relu	0.2	16	0.01
Opt, Desc. Vec.	5	A	-	-	-	-	-	700, relu	0	720, relu	0	600, relu	0.2	64	0.005
A1	1-5	B	Filters 32, Kernel 10	Max,5	Filters 32, Kernel 10	Max, 5	TRUE	20, relu	0.1	20, relu	0.1	-	-	32	0.001
A2	1-5	B	Filters 32, Kernel 10	Max,5	-	-	TRUE	20, relu	0.1	20, relu	0.1	-	-	32	0.001
A3	1-5	B	Filters 32, Kernel 10	Max,5	Filters 32, Kernel 10	Max, 5	TRUE	250, relu	0.1	250, relu	0.1	-	-	32	0.001
A4	1-5	B	Filters 32, Kernel 10	Max,5	-	-	TRUE	250, relu	0.1	250, relu	0.1	-	-	32	0.001
A5	1-5	B	Filters 32, Kernel 10	Max,5	Filters 32, Kernel 10	Max, 5	TRUE	500, relu	0.1	500, relu	0.1	-	-	32	0.001
A6	1-5	B	Filters 32, Kernel 10	Max,5	-	-	TRUE	500, relu	0.1	500, relu	0.1	-	-	32	0.001
Opt, OHE	1	B	Filters 24, Kernel 25	Avg, 6	-	-	TRUE	720, relu	0.3	-	-	-	-	32	0.001
Opt, OHE	2	B	Filters 16, Kernel 20	-	-	-	TRUE	620, relu	0.5	180, relu	0.2	-	-	32	0.001
Opt, OHE	3	B	Filters 32, Kernel 15	Avg, 3	-	-	TRUE	360, relu	0.2	420, relu	0	-	-	32	0.001
Opt, OHE	4	B	Filters 32, Kernel 20	Avg, 7	-	-	TRUE	360, relu	0	500, relu	0.5	-	-	32	0.005
Opt, OHE	5	B	Filters 24, Kernel 10	-	Filters 64, Kernel 5	Max, 7	TRUE	460, relu	0.3	-	-	-	-	64	0.005
Opt, Desc. Vec.	1	B	Filters 48, Kernel 5	-	-	-	TRUE	480, relu	0	660, relu	0	-	-	64	0.001
Opt, Desc. Vec.	2	B	Filters 56, Kernel 20	-	-	-	TRUE	580, relu	0.3	480, relu	0.1	-	-	256	0.001
Opt, Desc. Vec.	3	B	Filters 8, Kernel 20	-	-	-	TRUE	700, relu	0.3	740, relu	0	-	-	256	0.005
Opt, Desc. Vec.	4	B	Filters 56, Kernel 5	-	-	-	TRUE	460, relu	0	560, relu	0.3	-	-	256	0.001
Opt, Desc. Vec.	5	B	Filters 8, Kernel 20	Avg,6	-	Max, 6	TRUE	560, relu	0	420, relu	0.2	-	-	64	0.001

Table S1: This table lists the model hyperparameters corresponding to entries in Figure S1. The leftmost layers correspond to those appearing earlier in the model. Relevant hyperparameters associated with layers are provided in entries of the table. Entries with hyphens indicate the layer is not present in the model. For featurization strategies evaluated using hyperparameter-optimized models, each train-test split (five in total) results in a model with different hyperparameters. Distinctly, when evaluating featurization strategies with fixed architectures, the same hyperparameters are used to construct the model for all five splits.

S3 Description of Supporting Content

This section provides details of the other files present in the supporting content.

S3.1 Dataset Metadata

Metadata used for chemical unit representations specific to Datasets {A, B, C, D} are provided in csv titled Dataset{A,B,C,D}_metadata.csv. Each column can be taken as a different means of representation for the chemical unit of polymers in each dataset and were employed in scaled fingerprint and sequence explicit models.

S3.2 Dataset A Sequences and Labels

The sequences of intrinsically disordered proteins, obtained through the DISPROT database, that were modeled with the HPS molecular dynamics simulations are contained in Datdaset_A_Sequences.txt. Here, the residue identities are represented as numerical encodings. The identity of the residue corresponding to a particular number encoding can be found in DatasetA_metadata.csv. Their corresponding computed labels are provided in a csv file titled labels.csv.

S3.3 Model Error Metrics

Details on the model performances across all datasets is provided in a csv titled Model Performances.csv. Each line has a corresponding dataset, model / representation type, the identity of the chemical fingerprint used in the representation, strategy of handling degree of polymerization, and the indication of potential input and output transformations used. The remaining columns provide the MAE in absolute units and the goodness of fit R2 of the models, and their corresponding standard errors.