Supplementary Materials: PREFMoDeL: A systematic review and proposed taxonomy of biomolecular features for deep learning

Jacob L. North [1,2,3], Victor L. Hsu [4]

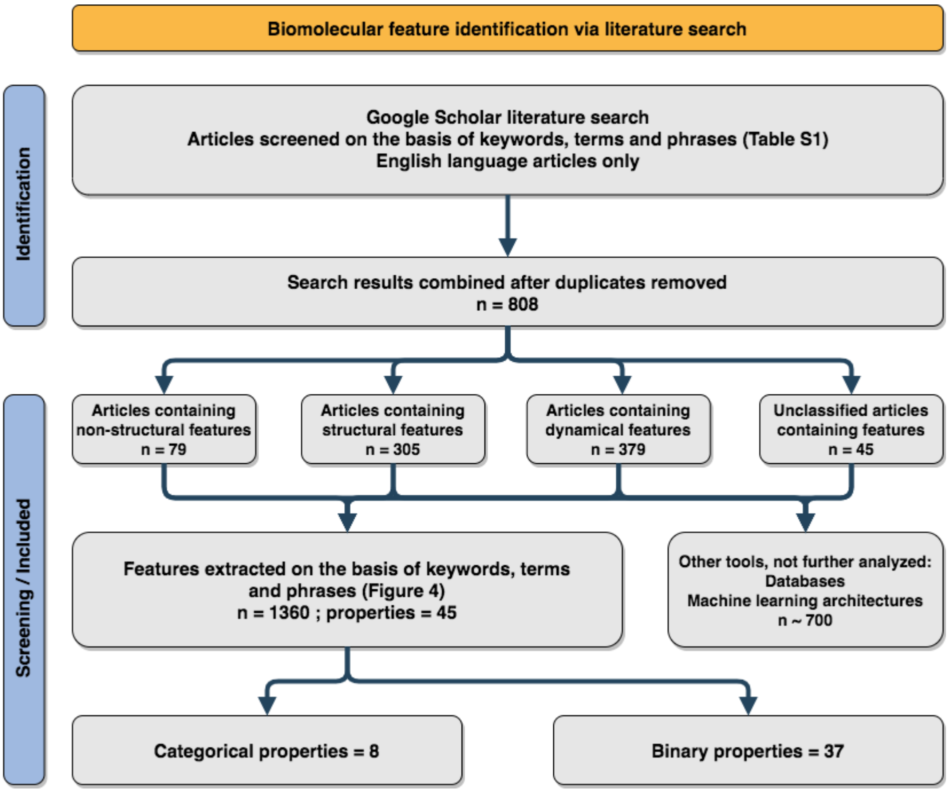
Biomolecular feature identification via literature search
Identification
Google Scholar literature search
Articles screened on the basis of keywords, terms and phrases (Table S1)
English language articles only
Search results combined after duplicates removed
n = 808
Screening / Included
Articles containing non-structural features
n = 79
Articles containing structural features
n = 305
Articles containing dynamical features
n = 379
Unclassified articles containing features
n = 45
Features extracted on the basis of keywords, terms and phrases (Figure 4)
n = 1360 ; properties = 45
Other tools, not further analyzed:
Databases
Machine learning architectures
n ~ 700
Categorical properties = 8
Binary properties = 37

Figure S1. The PRISMA 2020 flow diagram for the meta-analysis of biomolecular features.

```
$ cat patterns.txt
featur
CV
collective variable
represent
attribut
characterist
propert
model
```

Figure S2. Contents of patterns.txt, a plain text file containing one keyphrase per line used in parsing selected publication PDF files using `pdfgrep`. Words are incomplete so as to match a larger set of desired possible suffices.

Table S1. A table of mathematically-ideal properties and their significance.

Property	Significance	Example application
Representativeness (rep)	More representative features directly improve accuracy	All training tasks
Fixed dimension (fixed)	Network can accommodate the fixed dimension	Length of a molecular fingerprint
Continuity (cont)	Continuous predictions are valid	Most physical values of interest
Differentiability (diff)	Feature can be directly evaluated as a loss	Bond energy during minimization
Normalizability (norm)	Features are equally weighted	Values are properly scaled
Linearizability (lin)	Values can be compared across orders-of-magnitude	Values are properly scaled
Reversibility (rev)	Features can be trained in a different space	Cartesian density map in Fourier space
Example-uniqueness (uniq)	Featurization does not confuse data points	All training tasks
SE(3) invariance (SE3 invar)	Accommodates random orientations, less complexity, less training data	Ligand binding, protein structure prediction
SE(3) equivariance (SE3 equivar)	Accommodates random orientations, less complexity, less training data	Ligand binding, protein structure prediction

Table S2. Table of terms used in literature metasearch sorted by feature type.

Nonspecific	Drug design	Nonstructural	Structural	Dynamical
“feature engineering”	“enzyme design”	“graph”	“protein structure”	“dynamic”
“review”	“protein design”	“topology”	“structure”	“dynamical”
“machine learning”	“drug discovery”	“topological”	“structural”	“molecular dynamics”

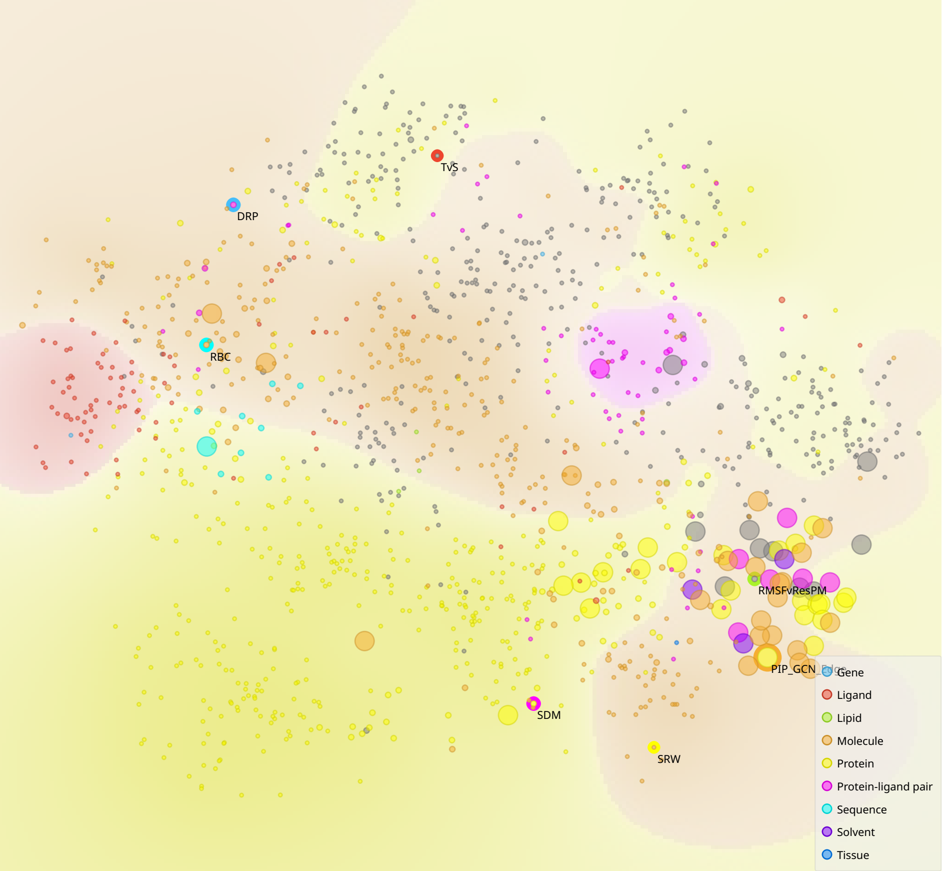

Figure S3. An example of feature clustering analysis with t-distributed stochastic neighbor embedding (t-SNE) where point color describes the biomolecule type (types indicated in the provided legend), point size reflects the ability to represent bonds (large points can represent bonds), and selected features are labeled by technique name. Seven unique features were selected, shown by a thick colored outline. From the top moving clockwise, TvS is temperature versus structural state; RMSFvResPM is the root-mean-square fluctuation versus the residue pairwise matrix; PIP_GCN_Edge is an edge feature vector for protein interface prediction; SRW is a spectral random walk; SDM is simulated electron density; RBC is the rotatable bond count; and DRP is a double reciprocal plot. Similarity is indicated by the Euclidean distance between features. Figure generated in Orange3.

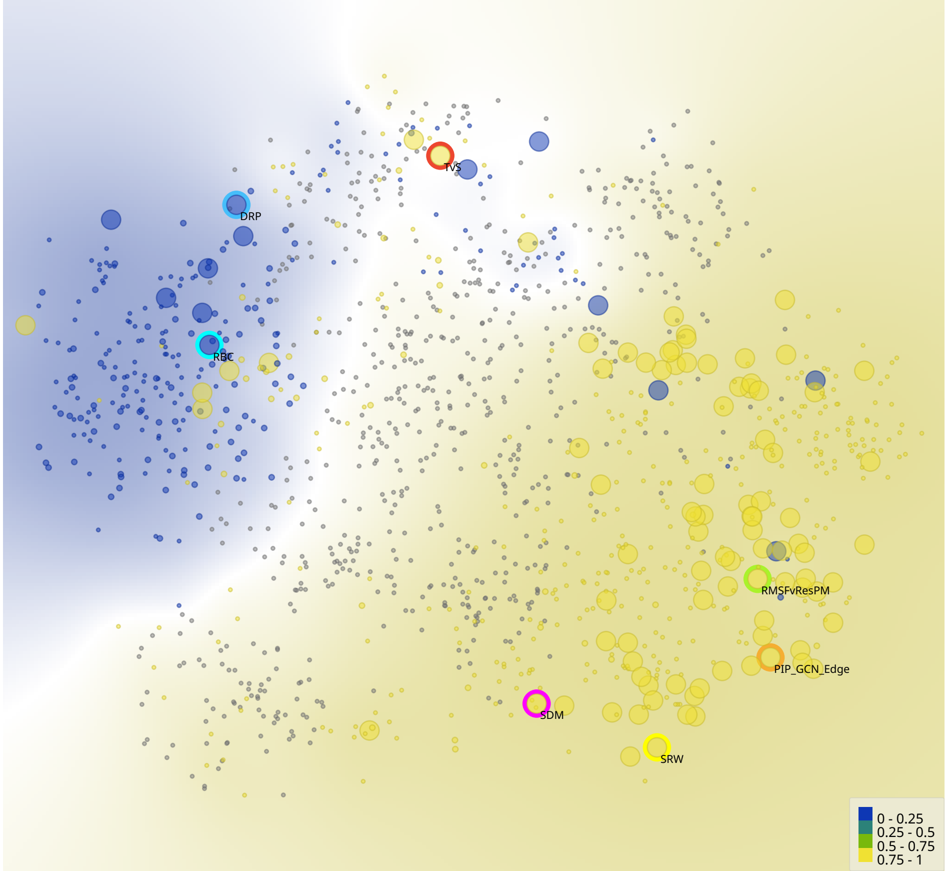

Figure S4. An example of feature clustering analysis with t-distributed stochastic neighbor embedding (t-SNE) where point color describes the ability to represent structure, point size describes the ability to represent molecular complexes (large points can represent complexes), and selected features are labeled by technique name. Seven unique features were selected, shown by a thick colored outline. From the top moving clockwise, TvS is temperature versus structural state; RMSFvResPM is the root-mean-square fluctuation versus the residue pairwise matrix; PIP_GCNEdge is an edge feature vector for protein interface prediction; SRW is a spectral random walk; SDM is simulated electron density; RBC is the rotatable bond count; and DRP is a double reciprocal plot. Similarity is indicated by the Euclidean distance between features. Figure generated in Orange3.

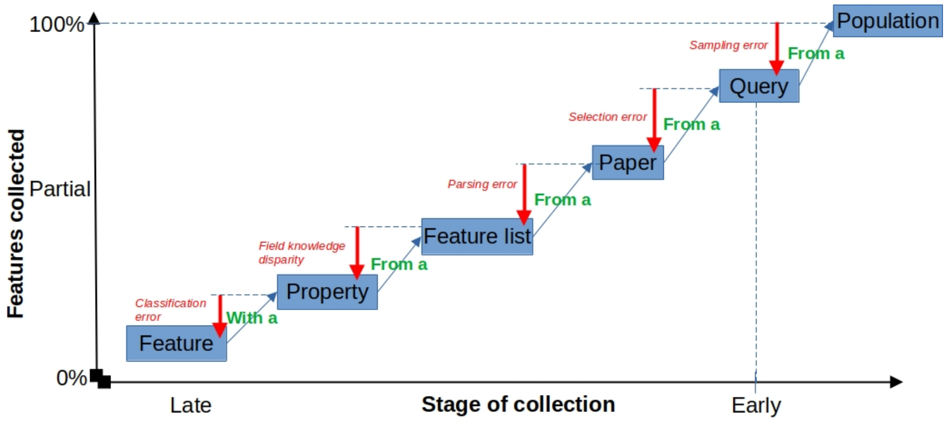

Figure S5. Conceptual relation of six types of data-containing objects used in feature collection, arranged by the number of features in a given paper they describe, their stage of collection, and the error incurred during collection. Feature rows are classified by their properties derived from the feature list which is constructed from reading a paper collected from a Google Scholar query from the population of research literature. Each step incurs a new form of error which reduces the proportion of total features collected. It is impractical to measure all features in the population of research literature. None of the objects in this diagram, aside from the population, are singular: multiple feature rows are described by multiple properties deriving from multiple feature lists, which are collected from multiple papers reflecting multiple Google Scholar queries from a single population of research literature.