Journal of Medicinal Chemistry

pubs.acs.org/jmc

Article

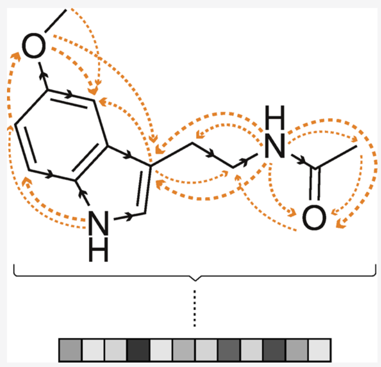
Improvement in ADMET Prediction with Multitask Deep Featurization

Evan N. Feinberg,* Elizabeth Joshi, Vijay S. Pande, and Alan C. Cheng*

Cite This: J. Med. Chem. 2020, 63, 8835–8848

Read Online

ABSTRACT: The absorption, distribution, metabolism, elimination, and toxicity (ADMET) properties of drug candidates are important for their efficacy and safety as therapeutics. Predicting ADMET properties has therefore been of great interest to the computational chemistry and medicinal chemistry communities in recent decades. Traditional cheminformatics approaches, using learners such as random forests and deep neural networks, leverage fingerprint feature representations of molecules. Here, we learn the features most relevant to each chemical task at hand by representing each molecule explicitly as a graph. By applying graph convolutions to this explicit molecular representation, we achieve, to our knowledge, unprecedented accuracy in prediction of ADMET properties. By challenging our methodology with rigorous cross-validation procedures and prognostic analyses, we show that deep featurization better enables molecular predictors to not only interpolate but also extrapolate to new regions of chemical space.



INTRODUCTION

The absorption, distribution, metabolism, elimination, and toxicity (ADMET) properties of drug candidates are important for their efficacy and safety as therapeutics. Historically, up to 50% of clinical trial failures have been attributed to deficiencies in ADMET properties.[1,2] Drug discovery teams have responded by working to reduce the failure rate with an increased focus on optimization of ADMET properties along with potency[3] and selectivity.[4]

Over the past few years, Merck & Co. has been heavily investing in leveraging institutional knowledge in an effort to drive hypothesis-driven, model-guided experimentation early in discovery. To that end, in silico models have been established for many of Merck's early screening assay end points deemed critical in the design of suitable potential candidates in order to selectively invest available resources in chemical matter having the best possible chance of delivering an efficacious and safe drug candidate in a timely fashion.[5,6]

Supervised machine learning (ML) is commonly used to model ADME end points and encompasses a family of functional forms and optimization schemes for mapping input features representing input samples to ground truth output labels. The traditional paradigm of ML involves representing molecules as one-dimensional vectors of features.[7] This *featurization* step frequently entails domain-specific knowledge.

In ligand-based QSAR, computational chemists have devised schemes to represent individual molecules as vectors of features. A popular representation uses circular fingerprints[8] of bond diameter 4 (i.e., ECFP4) to convert local neighborhoods of each atom to bits in a fixed vector. In contrast, atom pair features[9] denote pairs of atoms in a given molecule, the atom types, and the minimum bond path length that separates the two atoms. To perform ML, such fixed length vector featurizations will be paired with a learning algorithm of choice, such as random forest (RF), support vector machines, and multilayer perceptron deep neural networks (MLPs),[10] that will map the features to an output assay label of interest.

While circular fingerprints, atom pair features, MACCS keys,[11] and related generic schemes have been tremendously useful, they are inherently limited by the inefficiency of projecting a complex multidimensional object onto a single dimension. There is no meaning to proximity of bits along either an ECFP or pair descriptor set of molecular features. For instance, considering the pair descriptor framework, an sp^3 carbon that is five bonds away from an sp^2 nitrogen might denote the first bit in the feature vector, an sp^2 oxygen that is two bonds away from an sp^3 nitrogen might denote the very next bit in the same feature vector, and an sp^3 carbon that is four bonds away from an sp^2 nitrogen might denote the hundredth bit in the feature vector. Put in descriptor language, a feature like "CX3sp3-04-NX1sp2" is conceptually similar to "CX20sp3-04-NX2sp2", but the descriptors are treated as fully unrelated in descriptor-based QSAR. The arbitrary arrange-

Special Issue: Artificial Intelligence in Drug Discovery

Received: December 31, 2019
Published: April 14, 2020

https://dx.doi.org/10.1021/acs.jmedchem.9b02187
J. Med. Chem. 2020, 63, 8835–8848

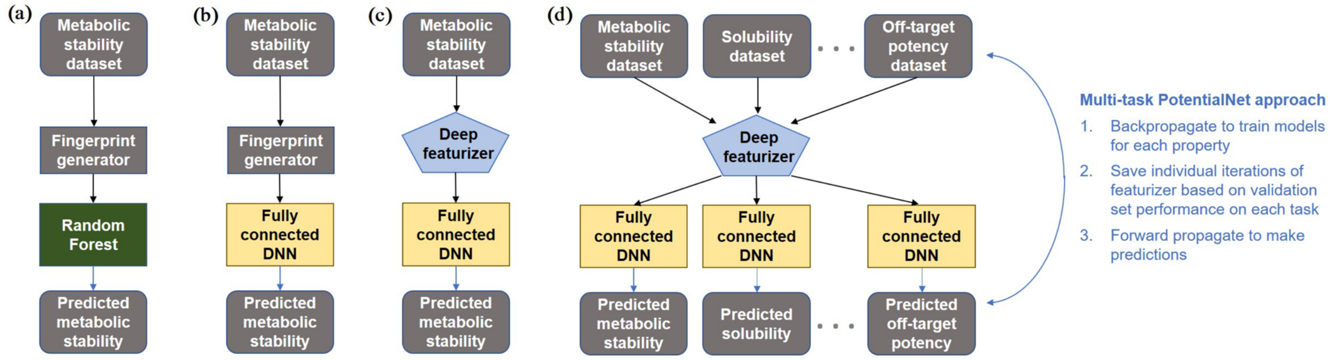


Figure 1. ML approaches. (a) Traditional ML uses a standard fingerprint generator and algorithms such as RF to learn a mapping from fixed fingerprint features to assay values. (b) MLPs replace the traditional algorithms with fully connected deep neural networks. (c) GCNNs such as PotentialNet learn features at the same time as learning the mapping to assay values. (d) The multitask GCNN (MT-PotentialNet) described here uses multiple assays to learn more generalized features at the same time as learning the mapping to assay values.

ment of bits in the featurization process must be "relearned" by the ML algorithm.

On the other hand, graph convolutional approaches can separate the "element" component from the "hybridization" component and both from the "bond distance" component. Graph convoluions thus can represent the conceptual proximity between qualitatively similar descriptors. Just as two-dimensional convolutions win with images by exploiting pixel adjacency, and recurrent neural networks win with speech translation by exploiting temporal adjacency, graph convolutions can win with molecules by exploiting the concept of atom and bond adjacency.

The contrast between MLPs and a graph convolutional neural networks (GCNNs) for chemical machine learning is illustrated in Figure 6. With MLPs, each molecule is represented by a vector, whereas with GCNNs, each molecule is represented by both an adjacency matrix and a feature matrix. The GCNN uses multiple graph convolutional layers (blue pentagons in Figure 1) that generate an end-to-end differentiable fingerprint vector. The final layers of the GCNN are identical in form to the hidden layers (yellow boxes in Figure 1) of the MLP. The GCNN learns a feature vector in the graph convolution layers. Additional details are presented in Methods.

Since the advent of the graph convolutional neural network to chemistry ML,[12] a spate of new approaches[13−17] have improved upon the basic graph convolutional layers expressed in Figure 1. Here, we train a derivative of PotentialNet GCNNs,[17] which have several key differences from the earliest graph neural networks (cf. Methods for more details). Whereas one can view the basic GCNN (Figure 6) as analogous to a learnable and more efficient ECFP featurization,[12] one can view the PotentialNet graph layers as analogous to a learnable and more efficient pair descriptor[9] featurization. The second advance is the use of multitask featurization (Figure 1). Like multitask learning, multitask featurization enables PotentialNet (in this variant, MT-PotentialNet) to learn features from multiple assays, leading to greater performance and generalization, and can be compared to a medicinal chemist learning patterns across multiple projects.

In this paper, we conduct a direct comparison of the current state-of-the-art algorithm similar to those used by many major pharmaceutical companies (random forests based on atom pair descriptors) with GCNNs (single-task and multitask PotentialNet, see Methods). We trained ML models on 31 chemical data sets describing results of various ADMET assays (ranging from physicochemical properties to *in vivo* PK properties) and compared results of random forests with those of GCNNs on held-out test sets chosen by two different cross-validation strategies (further details on model training are included in Methods). In addition, we ascertain the capacity of our trained models to generalize to assays conducted outside our institution by downloading data sets from the scholarly literature and conducting further performance comparisons. Finally, we conduct a prognostic comparison of prediction accuracy of RF and PotentialNet GCNNs on assay data of new chemical entities recorded well after all model parameters were frozen in place.

RESULTS

The primary purpose of supervised machine learning is to train computer models to make accurate predictions about samples that have not been seen before in the learning process. In the discipline of computer vision, the ability to interpolate between training samples is often sufficient for real-world applications. Such is not the case in the field of medicinal chemistry. When a chemist is tasked with generating new molecular entities to selectively modulate a given biological target, they must invent chemical matter that is fundamentally different from previously known materials. This need stems from both scientific and practical concerns; every biological target is different and likely requires heretofore nonexistent chemical matter, and the patent system demands that, to garner protection, new chemical entities must be not only useful but also novel and sufficiently different from currently existing molecules.

Cross-validation is a subtle yet critical component of any ML initiative. In the absence of the ability to gather prospective data, it is standard practice in ML to divide one's retrospectively available training data into three disjoint subsets: train, valid, and test (though it is only strictly necessary that the test set be disjoint from the others). It is well-known that cross-validation strategies typically used in the vision or natural language domains, like random splitting, significantly overestimate the generalization and extrapolation ability of machine learning methods.[18] As a result, we deploy two train−test splits that, compared to random splitting, are at once more challenging and also more accurately reflect the real world task of drug discovery. First, we split all data sets temporally, training on molecules assayed before the earliest $date_i$, selecting models based on performance on molecules

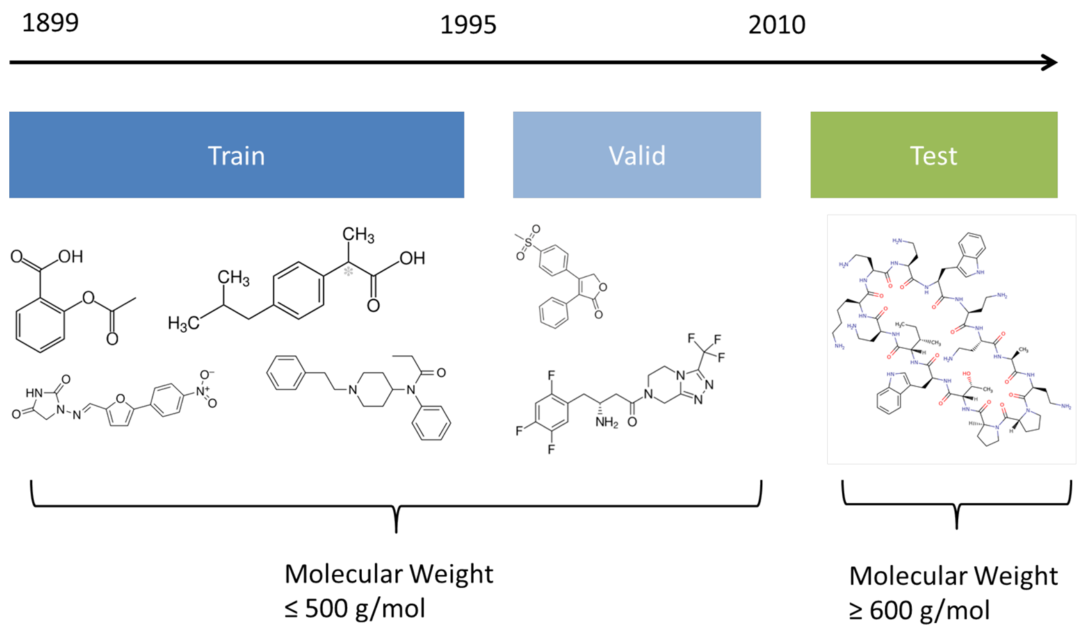


Figure 2. Temporal plus molecular weight split. In this rigorous cross-validation procedure, one trains and selects models based on older, smaller molecules, with the resultant best model (according to the validation set score) being evaluated a single time on a held-out test set on newer, larger molecules. Specifically, the train and validation sets comprise only molecules below a certain molecular weight threshold and synthesized before a certain date, whereas the test set comprises only molecules above a certain molecular weight threshold and synthesized after a certain date.

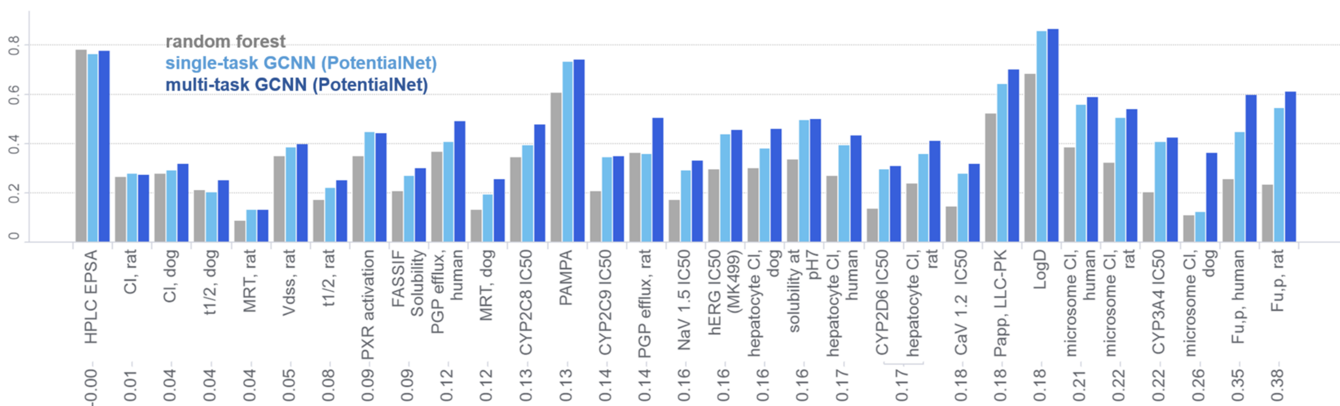


Figure 3. Temporal split: performance of PotentialNet versus RF for all assays.

assayed between $date_i$ and intermediate $date_j$, and evaluating the final model on held-out molecules assayed after the latest $date_j$. Such temporal splitting is meant to parallel the types of progressions that typically occur in lead optimization programs as well as reflect broader changes in the medicinal chemistry field.

In addition to temporal splitting, we introduce an additional cross-validation strategy in which we *both* divide train, valid, and test sets temporally *and* add the following challenge: (1) removal of molecules with molecular weight greater than 500 g/mol from the training and validation sets and (2) inclusion of only molecules with molecular weight greater than or equal to 600 g/mol in the test set. We denote this as *temporal plus molecular weight* split (Figure 2). We chose to include this additional cross-validation split to provide a more challenging estimate of the ability of the respective ML models to extrapolate in chemical space. In addition, as macrocycles and other medium-sized molecules become more prominent in drug discovery pipelines, we investigated the applicability of ML models trained on current, small molecule-focused data sets to these larger MW modalities.

Temporal Split. In aggregate, PotentialNet achieves a 64% average improvement and a 52% median improvement in r^2 over RF across all 31 reported data sets (Figure 3, Table S1, Table 1). The mean r^2 over the various test data sets is 0.30 for RF and 0.44 for MT-PotentialNet, corresponding to a mean $\overline{\Delta r^2}$(MT-PotentialNet − RF) = 0.15. Among the assays for which MT-PotentialNet offers the most improvement (Figure 10) are plasma protein binding (fraction unbound for both human, Δr^2 = 0.34, and rat, Δr^2 = 0.38), microsomal clearance (human, Δr^2 = 0.21; dog, Δr^2 = 0.26; rat, Δr^2 = 0.22), CYP3A4 inhibition (Δr^2 = 0.22), log *D* (Δr^2 = 0.18), and passive membrane absorption (Δr^2 = 0.18). Meanwhile, assays HPLC EPSA and rat clearance, which account for 2 of the 31 assays, show no statistically significant difference. All r^2 values are reported with confidence intervals computed according to ref 19.

In addition to improvements in r^2, for both cross-validation schemes discussed, we note that the slope of the linear

Table 1. Performance: Temporal Split

data set	RandomForest R^2	RandomForest R^2, 95% CI	PotentialNet R^2	PotentialNet R^2, 95% CI	absolute improvement	percentage improvement
HPLC EPSA	0.775	(0.76, 0.789)	0.772	(0.757, 0.786)	−0.003	−0.340
Cl, rat	0.260	(0.25, 0.271)	0.272	(0.262, 0.282)	0.011	4.360
Cl, dog	0.277	(0.258, 0.295)	0.314	(0.296, 0.333)	0.038	13.579
$t_{1/2}$, dog	0.209	(0.192, 0.225)	0.247	(0.23, 0.264)	0.038	18.455
MRT, rat (h)	0.084	(0.075, 0.093)	0.127	(0.117, 0.138)	0.044	52.333
Vd_{ss}, rat (L/kg)	0.345	(0.335, 0.356)	0.393	(0.383, 0.403)	0.048	13.769
$t_{1/2}$, rat	0.168	(0.16, 0.177)	0.248	(0.239, 0.258)	0.080	47.625
PXR activation	0.348	(0.342, 0.353)	0.438	(0.432, 0.443)	0.090	25.933
solubility in FASSIF	0.203	(0.197, 0.208)	0.296	(0.291, 0.302)	0.094	46.178
PGP efflux, human	0.366	(0.35, 0.382)	0.489	(0.474, 0.503)	0.123	33.572
MRT, dog (h)	0.129	(0.112, 0.147)	0.253	(0.231, 0.274)	0.124	96.050
CYP2C8 IC_{50}	0.343	(0.331, 0.355)	0.474	(0.462, 0.485)	0.130	38.029
PAMPA	0.604	(0.553, 0.651)	0.735	(0.697, 0.77)	0.131	21.752
CYP2C9 IC_{50}	0.205	(0.2, 0.21)	0.346	(0.341, 0.352)	0.141	68.728
PGP efflux, rat	0.358	(0.341, 0.374)	0.502	(0.487, 0.517)	0.145	40.497
NaV 1.5 IC_{50}	0.170	(0.164, 0.177)	0.328	(0.321, 0.336)	0.158	92.688
hERG inh (MK499)	0.291	(0.286, 0.295)	0.451	(0.447, 0.456)	0.161	55.306
hepatocyte Cl, dog	0.298	(0.266, 0.329)	0.459	(0.429, 0.489)	0.162	54.326
solubility at pH 7	0.331	(0.327, 0.335)	0.496	(0.493, 0.5)	0.165	49.843
hepatocyte Cl, human	0.265	(0.252, 0.279)	0.431	(0.417, 0.444)	0.165	62.323
CYP2D6 IC_{50}	0.135	(0.131, 0.14)	0.306	(0.3, 0.312)	0.171	126.063
hepatocyte Cl, rat	0.237	(0.223, 0.251)	0.408	(0.394, 0.422)	0.171	71.951
CaV 1.2 IC_{50}	0.141	(0.135, 0.147)	0.316	(0.31, 0.323)	0.175	124.278
P_{app}, LLC-PK	0.520	(0.509, 0.532)	0.696	(0.688, 0.704)	0.176	33.781
HPLC log *D*	0.678	(0.67, 0.686)	0.861	(0.857, 0.865)	0.183	26.984
microsome Cl, human	0.382	(0.369, 0.394)	0.588	(0.578, 0.598)	0.206	53.992
microsome Cl, rat	0.320	(0.307, 0.333)	0.539	(0.528, 0.551)	0.219	68.500
CYP3A4 IC_{50}	0.200	(0.194, 0.205)	0.420	(0.415, 0.426)	0.221	110.634
microsome Cl, dog	0.105	(0.076, 0.137)	0.361	(0.32, 0.402)	0.257	245.523
Fu,p human	0.251	(0.233, 0.27)	0.596	(0.58, 0.611)	0.344	136.910
Fu,p rat	0.233	(0.221, 0.245)	0.608	(0.598, 0.618)	0.375	161.049
median	0.265		0.431		0.158	52.3
mean	0.300		0.444		0.147	64.34

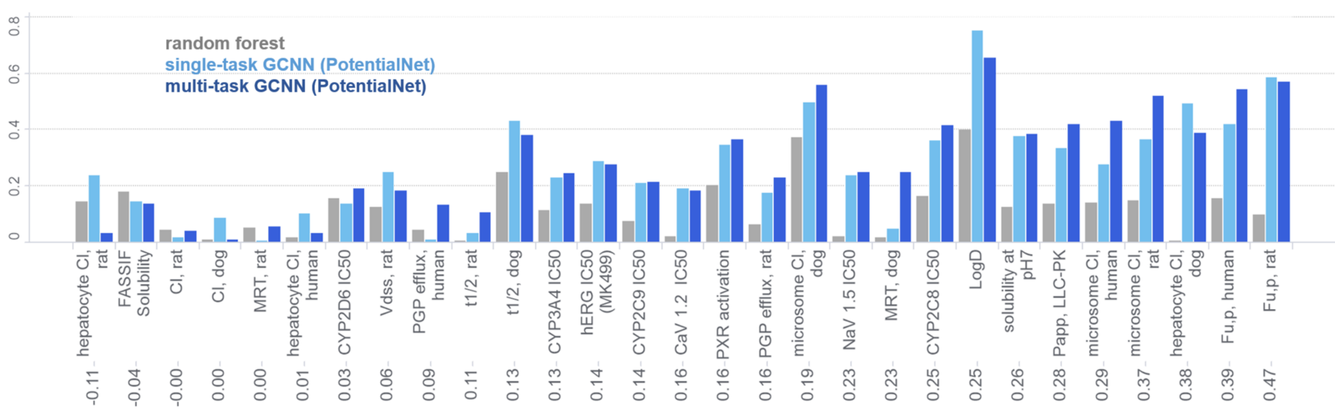


Figure 4. Temporal plus molecular weight split: performance of PotentialNet versus RF for all assays.

regression line between predicted and experimental data is, on average, closer to unity for MT-PotentialNet than it is for RF. This difference is qualitatively notable in Figure 10. A corollary, which is also illustrated by Figures 10, 11, and 12, is that MT-PotentialNet DNNs perform noticeably better than RF in predicting the correct *range* of values for a given prediction task. At Merck & Co., this deficiency of RF is in part rectified by *ex post facto* prediction rescaling, which in part recovers the slope but makes no difference in r^2.

The commercially available molecules for which MT-PotentialNet achieved the greatest improvement in prediction versus RF are displayed in Table S12. An example of a molecule on which RF renders a more accurate prediction is shown in Table S13.

Temporal plus Molecular Weight Split. We then investigated the cross-validation setting where data were both (a) split temporally and (b) molecules with MW > 500 g/mol were removed from the training set while *only* molecules with

https://dx.doi.org/10.1021/acs.jmedchem.9b02187
J. Med. Chem. 2020, 63, 8835−8848

Table 2. Performance: Temporal plus Molecular Weight Split

data set	RandomForest R^2	RandomForest R^2, 95% CI	PotentialNet R^2	PotentialNet R^2, 95% CI	absolute improvement	percentage improvement
hepatocyte Cl, rat	0.142	(0.087, 0.206)	0.030	(0.007, 0.068)	−0.112	−78.891
solubility in FASSIF	0.176	(0.158, 0.194)	0.136	(0.12, 0.153)	−0.040	−22.804
Cl, rat	0.041	(0.025, 0.061)	0.039	(0.023, 0.059)	−0.002	−5.284
Cl, dog	0.008	(0.0, 0.037)	0.008	(0.0, 0.037)	−0.000	−0.281
MRT, rat (h)	0.049	(0.028, 0.074)	0.054	(0.032, 0.08)	0.005	10.590
hepatocyte Cl, human	0.017	(0.003, 0.044)	0.029	(0.008, 0.062)	0.012	67.114
CYP2D6 IC_{50}	0.154	(0.134, 0.175)	0.188	(0.166, 0.21)	0.034	21.943
Vd_{ss}, rat (L/kg)	0.122	(0.095, 0.153)	0.182	(0.15, 0.216)	0.060	48.745
PGP efflux, human	0.043	(0.008, 0.101)	0.133	(0.066, 0.215)	0.091	212.901
$t_{1/2}$, rat	0.000	(0.001, 0.004)	0.106	(0.081, 0.134)	0.106	23708.598
$t_{1/2}$, dog	0.247	(0.176, 0.321)	0.377	(0.302, 0.451)	0.131	52.942
CYP3A4 IC_{50}	0.111	(0.093, 0.129)	0.242	(0.219, 0.265)	0.131	118.507
CYP2C9 IC_{50}	0.075	(0.061, 0.092)	0.214	(0.192, 0.236)	0.138	182.877
hERG inh (MK499)	0.135	(0.123, 0.148)	0.273	(0.258, 0.289)	0.138	102.122
CaV 1.2 IC_{50}	0.021	(0.013, 0.031)	0.181	(0.159, 0.204)	0.160	767.023
PXR activation	0.201	(0.18, 0.223)	0.364	(0.341, 0.387)	0.163	80.830
PGP efflux, rat	0.063	(0.018, 0.132)	0.226	(0.14, 0.32)	0.163	256.734
microsome Cl, dog	0.371	(0.193, 0.544)	0.556	(0.385, 0.695)	0.185	49.751
NaV 1.5 IC_{50}	0.018	(0.01, 0.028)	0.247	(0.222, 0.273)	0.229	1279.873
MRT, dog (h)	0.016	(0.0, 0.062)	0.249	(0.159, 0.346)	0.233	1416.502
CYP2C8 IC_{50}	0.163	(0.116, 0.213)	0.412	(0.357, 0.466)	0.250	153.437
HPLC log D	0.397	(0.354, 0.439)	0.651	(0.618, 0.682)	0.254	63.935
solubility at pH 7	0.124	(0.11, 0.138)	0.384	(0.367, 0.402)	0.261	210.535
P_{app}, LLC-PK	0.137	(0.088, 0.194)	0.418	(0.354, 0.479)	0.280	204.100
microsome Cl, human	0.139	(0.089, 0.196)	0.430	(0.366, 0.491)	0.291	208.433
microsome Cl, rat	0.146	(0.094, 0.205)	0.519	(0.457, 0.577)	0.373	255.865
hepatocyte Cl, dog	0.003	(0.012, 0.047)	0.387	(0.259, 0.509)	0.384	12724.229
Fu,p human	0.154	(0.11, 0.202)	0.542	(0.493, 0.587)	0.387	251.270
Fu,p rat	0.097	(0.065, 0.134)	0.567	(0.526, 0.606)	0.470	486.274
median	0.122		0.247		0.160	153.4
mean	0.116		0.281		0.165	1476.8

MW > 600 g/mol were retained in the test set (Figure 2). In aggregate, MT-PotentialNet achieves a 153% median improvement in r^2 over RF across all 29 reported data sets. PAMPA and EPSA are not included due to insufficient number of compounds meeting the training and testing criteria (Table S2, Figure 4, Table 2). The mean r^2 over the various test data sets is 0.12 for RF and 0.28 for MT-PotentialNet, corresponding to a mean Δr^2(MT-PotentialNet − RF) = 0.16. The assays for which MT-PotentialNet offers the most improvement (Figure 11) are plasma protein binding (fraction unbound for both human, $\Delta r^2 = 0.38$, and rat, $\Delta r^2 = 0.47$), microsomal clearance (human, $\Delta r^2 = 0.29$; dog, $\Delta r^2 = 0.19$; rat, $\Delta r^2 = 0.37$), CYP2C8 inhibition ($\Delta r^2 = 0.25$), log D ($\Delta r^2 = 0.25$), and passive membrane absorption ($\Delta r^2 = 0.28$). Meanwhile, human hepatocyte clearance, CYP2D6 inhibition, rat and dog clearance, dog half-life, human PGP (efflux), rat MRT, and rat volume of distribution exhibit no statistically significant difference in model predictivity ($\frac{8}{29}$ of all data sets), with only rat hepatocyte clearance being predicted less well for MT-PotentialNet as compared to RF. It should be noted that the quantity of molecules in the test sets are smaller in Temporal plus MW split as compared to temporal only split, and therefore, it is accordingly more difficult to reach statistically significant differences in r^2 (Supporting Information Table S7).

The commercially available molecules for which MT-PotentialNet achieved the greatest improvement in prediction versus RF are displayed in Table S14. It is intriguing that the same molecule undergoes the greatest improvement for both human fraction unbound as well as CYP2D6 inhibition.

As previous works have noted,[20] multitask style training can boost (or, less charitably, inflate) the performance of neural networks by sharing information between the training molecules of one task and the test molecules of another task, especially if the activities are in some way correlated. Another advantage of the Temporal plus MW cross-validation approach is that it mitigates hemorrhaging of information between assay data sets. By introducing a minimum 100 g/mol MW gap between train and test molecules, it not only is impossible for train molecules in one task to appear as test molecules for another task but also circumscribes the similarity of any given task's training set to any given other task's test set. We further investigate the relative effect of multitask versus single task training in Supporting Information Table S5. However, even in cases where there is similarity or even identity between training molecules of one assay and test molecules of another assay, in the practice of chemical machine learning, this may in fact be desirable in cases. For instance, if less expensive, cell-free assays, like solubility, have a strong correlation with a more expensive end point, like dog mean residence time, it would be an attractive property of multitask learning if solubility data on molecules in a preexisting database could inform more accurate predictions of the animal mean residence time of untested, similar molecules.

https://dx.doi.org/10.1021/acs.jmedchem.9b02187
J. Med. Chem. 2020, 63, 8835−8848

Table 3. Performance: Prospective Split

data set	RandomForest R^2	RandomForest R^2, 95% CI	PotentialNet R^2	PotentialNet R^2, 95% CI	absolute improvement	percentage improvement
Cl, dog	0.231	(0.163, 0.304)	0.173	(0.111, 0.242)	−0.058	−25.162
HPLC EPSA	0.819	(0.806, 0.831)	0.781	(0.765, 0.795)	−0.039	−4.710
Cl, rat	0.350	(0.317, 0.384)	0.330	(0.297, 0.363)	−0.020	−5.770
$t_{1/2}$, dog	0.326	(0.257, 0.396)	0.326	(0.257, 0.395)	−0.001	−0.162
$t_{1/2}$, rat	0.318	(0.285, 0.351)	0.319	(0.286, 0.352)	0.001	0.215
Vd_{ss}, rat (L/kg)	0.520	(0.49, 0.55)	0.525	(0.494, 0.554)	0.004	0.846
microsome Cl, human	0.542	(0.503, 0.579)	0.579	(0.542, 0.614)	0.037	6.878
microsome Cl, rat	0.478	(0.436, 0.519)	0.541	(0.501, 0.579)	0.063	13.069
CYP2C9 IC_{50}	0.257	(0.232, 0.282)	0.322	(0.296, 0.347)	0.065	25.396
CaV 1.2 IC_{50}	0.120	(0.103, 0.138)	0.192	(0.172, 0.213)	0.072	59.540
NaV 1.5 IC_{50}	0.137	(0.119, 0.155)	0.213	(0.192, 0.234)	0.076	55.582
solubility in FASSIF	0.199	(0.188, 0.21)	0.286	(0.274, 0.298)	0.087	43.823
CYP3A4 IC_{50}	0.255	(0.236, 0.274)	0.344	(0.325, 0.364)	0.089	35.108
PAMPA	0.378	(0.305, 0.45)	0.474	(0.403, 0.541)	0.096	25.487
PXR activation	0.325	(0.308, 0.342)	0.431	(0.414, 0.448)	0.106	32.608
PGP efflux, human	0.349	(0.252, 0.446)	0.468	(0.372, 0.557)	0.119	33.934
CYP2D6 IC_{50}	0.287	(0.261, 0.313)	0.406	(0.38, 0.432)	0.120	41.735
P_{app}, LLC-PK	0.111	(0.066, 0.164)	0.241	(0.18, 0.304)	0.129	116.526
hERG inh (MK499)	0.306	(0.286, 0.325)	0.445	(0.425, 0.464)	0.139	45.555
hepatocyte Cl, rat	0.195	(0.161, 0.231)	0.342	(0.305, 0.38)	0.147	75.254
Fu,p rat	0.481	(0.45, 0.511)	0.642	(0.617, 0.666)	0.161	33.522
hepatocyte Cl, dog	0.224	(0.15, 0.304)	0.407	(0.326, 0.486)	0.183	81.705
PGP efflux, rat	0.293	(0.076, 0.532)	0.477	(0.234, 0.681)	0.185	63.129
hepatocyte Cl, human	0.255	(0.22, 0.29)	0.445	(0.411, 0.479)	0.190	74.741
microsome Cl, dog	0.518	(0.136, 0.794)	0.739	(0.42, 0.899)	0.221	42.766
HPLC log *D*	0.402	(0.39, 0.413)	0.705	(0.697, 0.712)	0.303	75.503
Fu,p human	0.582	(0.3, 0.78)	0.919	(0.832, 0.962)	0.337	57.903
median	0.318		0.431		0.096	35.1
mean	0.343		0.447		0.104	37.2

Held-Out Data from Literature. To further ascertain the generalization capacity of our models, we obtained data from scholarly literature. In particular, we obtained data on macrocyclic compounds for passive membrane permeability and log *D* from ref 21, a study selected for its comparatively large number of measurements on high molecular weight species measured with a consistent assay. We observed a statistically significant increase in performance (Table 4) for both passive membrane permeability,

$$\Delta r^2(\text{MT-PotentialNet} - \text{RF}) = 0.23$$

and log *D*,

$$\Delta r^2(\text{MT-PotentialNet} - \text{RF}) = 0.21$$

The four molecules for which MT-PotentialNet exhibits the greatest improvement in predictive accuracy over RF are shown in Table S15.

The second molecule in Table S15 is experimentally quite permeable, which MT-PotentialNet correctly identifies but RF severely underestimates. Note that the aliphatic tertiary amine would likely be protonated and therefore charged at physiologic pH. The proximity of an ether oxygen may "protect" the charge, increasing the ability to passively diffuse through lipid bilayers. Because of the relative efficiency with which information traverses bonds in a graph convolution as opposed to the fixed pair features that are provided to the random forest, it is intuitively straightforward for a graph neural network to learn the "atom type" of a high pK_a nitrogen in spatial proximity to an electron rich oxygen, whereas pair features would rigidly specify an aliphatic nitrogen three bonds away from an aliphatic oxygen.

Study on Subsequent Data. In September 2018, we froze the parameters of RF and MT-PotentialNet models trained on all available assay data recorded internally at Merck & Co. up to the end of August 2018. After approximately two months had elapsed after the registration of the last training data point, we evaluated the performance of those *a priori* frozen models on new experimental assay data gathered on compounds registered between November 2018 and the end of February 2019. This constitutes a rigorous, forward-looking analysis in which there is a two month gap between training data collection and model evaluation of future data, further challenging the generalization capacity of the trained models. For statistical power, we chose to evaluate performance on all assays for which at least 10 compounds were experimentally tested in the period Nov 2018 to Feb 2019. Over these 27 assays, RF achieved a median r^2 of 0.32, whereas MT-PotentialNet achieved a median r^2 of 0.43 for a median Δr^2 = 0.10 (Table S3). Performance of each assay can be found in Supporting Information Figure S6 and Table 3, and scatter plots of predicted versus experimental values for several assays can be found in Figures 12 and 13. While it makes no

Table 4. Performance on Held-Out Data from Literature[21]

property	RandomForest r^2	PotentialNet r^2
P_{app}	0.150 (0.081, 0.232)	0.381 (0.292, 0.468)
log *D*	0.394 (0.305, 0.480)	0.603 (0.528, 0.670)

https://dx.doi.org/10.1021/acs.jmedchem.9b02187
J. Med. Chem. 2020, 63, 8835−8848

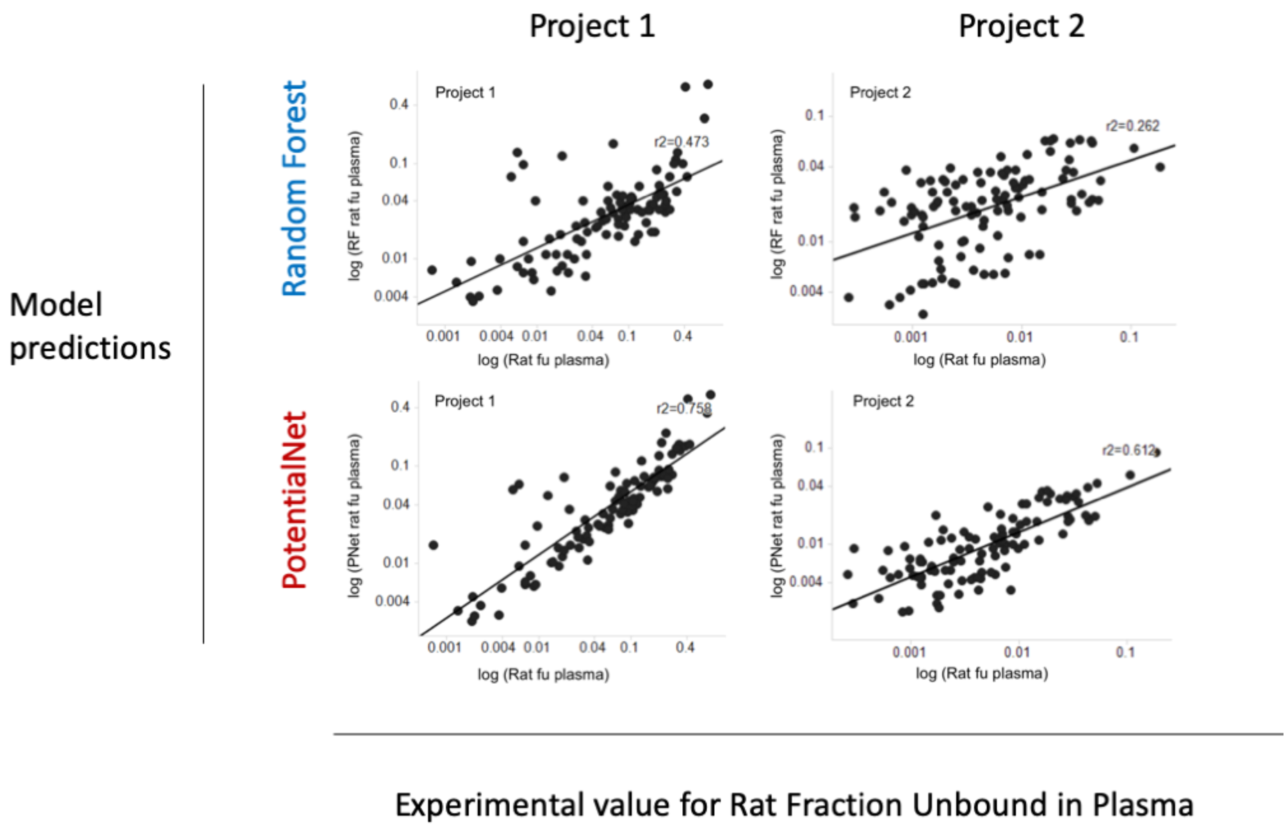


Figure 5. Prospective prediction of rat fraction unbound in plasma (rat fu,p) in two active projects using random forest models (top row) and PotentialNet models (bottom row) for compounds experimentally tested between September and December 2018. Data for project 1 and project 2 are for 97 and 123 compounds, respectively.

difference in r^2, we have chosen to scale the values predicted by both RF and by MT-PotentialNet to match the mean and standard deviation of the distribution of assay data in the training set to more faithfully reflect how these models would be used practically in an active pharmaceutical project setting.

Performance on Two Specific Projects. To assess performance on individual projects, we applied the August 2018 models to prediction of rat plasma fraction unbound on two currently active lead optimization projects at Merck & Co. The results (Figure 5) suggest that the performance on the individual projects are similar to the Temporal Split and Temporal plus MW split results.

Algorithm Ablation Study. Historically, as a discipline, machine learning arose from statistical learning, and a key line of inquiry in statistics involves extricating potentially confounding variables. Compared with RF, we introduce several algorithmic changes at once: use of neural network instead of random forest; use of graph convolution as a neural network architecture based on a graph adjacency and feature matrices as input rather than either RF or MLP based on flat one-dimensional features; and use of a variant of multitask learning rather than single task learning. How much of the performance gain accrued by PotentialNet can be attributed to each of the three changes? To investigate, we conducted an algorithm ablation study to compare performance contributions (Supporting Information Tables S5 and S9; we also include xgboost for additional comparison). It is reasonable to contend that one should solely compare RF with single task neural networks since the former is incapable of jointly learning on several assay data sets simultaneously. However, one of the intuitive advantages of a GCNN over *either* RF or MLP is that a GCNN can learn the atomic interaction features relevant to the prediction task at hand. Not only can graph convolutions learn the features, but adding different molecules from different tasks allows networks to learn, potentially with greater generalization, both through the effect of task correlation and learning richer features by incorporating a greater area of chemical space.

■ DISCUSSION AND CONCLUSIONS

Preclinical drug discovery is a critical, and often rate-limiting stage of the broader pharmaceutical development process. The multiobjective optimization among potency and ADMET properties, which can entail vexing trade-offs, is a critical bottleneck in preclinical discovery.[22,23] More accurate prediction of ADMET end points can both prevent exploration of undesirable chemical space and facilitate access to desirable regions of chemical space, thereby making preclinical discovery not only more efficient but perhaps more productive as well.

To assess if a modern graph convolutional neural network[17] succeeds in more accurately predicting ADMET end points, we conducted a rigorous performance comparison between our multitask GCNN, single-task GCNN, and a state-of-the-art random forest based on cheminformatic features. We included a total of 31 assay data sets in our analysis, employed two cross-validation splits (*temporal split*[18] and a combined *temporal plus molecular weight split*), and made predictions on a publicly available held-out test set. Finally, we compare predictions on future data made with RF, single-task PotentialNet GCNN, and multitask PotentialNet GCNN.

https://dx.doi.org/10.1021/acs.jmedchem.9b02187
J. Med. Chem. 2020, 63, 8835−8848

Encouragingly, statistical improvements were observed in each of the four validation settings; multitask GCNN shows improvements over the single-task GCNN and in turn over RF. In the temporal split setting, across 31 tasks, RF achieved a mean r^2 of 0.30, whereas MT-PotentialNet achieved a mean r^2 of 0.44 and single-task PotentialNet achieved a mean r^2 of 0.39 (Table S1). In the Temporal plus MW split setting (where only older smaller molecules were included in the training set while only newer larger molecules included in the test set) across 29 tasks, RF achieved a mean r^2 of 0.12, whereas our multitask GCNN achieved a mean r^2 of 0.28 and our single-task GCNN achieved a mean r^2 of 0.26 (Table S2). In the final pseudoprospective validation setting, we assessed the ability of pretrained RF and pretrained MT-PotentialNet models to predict passive membrane permeability and log D on an experimental data set on macrocycles obtained from the literature.[21] In this setting, for passive membrane permeability, RF models achieved an r^2 of 0.15 whereas our multitask GCNN achieved an r^2 of 0.38; for log D, RF achieved an r^2 of 0.39 whereas our multitask GCNN achieved an r^2 of 0.60 (Table S14).

While the three described retrospective investigations are more rigorous than random splitting and are meant to more faithfully reflect the generalization capacity of a model in the practical real world of pharmaceutical chemistry, we also believe that prospective validation is important whenever the resources are available to do so. To this end, we made predictions on 23 assays, each of which contained measurements for new chemical entities synthesized and evaluated after November 2018 (the last data point for model training was collected in August 2018). In aggregate, there is a mean Δr^2 of 0.10 for MT-PotentialNet over RF models (Figures 7, 8, and 9). This improvement in accuracy in a future and relatively constrained time window is largely consistent with that prognosticated by the retrospective temporal split study and is encouraging for the utility of deep featurization in a predictive capacity for drug discovery.

Multi-Layer Perceptron / Fully Connected Neural Network

$$h^{(1)} = ReLU\left(W^{(1)} \cdot x\right)$$
$$h^{(2)} = ReLU\left(W^{(2)} \cdot h^{(1)}\right)$$
$$\vdots$$
$$h^{(K)} = ReLU\left(W^{(K)} \cdot h^{(K-1)}\right)$$

Graph Convolutional Neural Network (GCNN)

$$H^{(1)} = ReLU\left(W^{(1)} \cdot A \cdot X\right)$$
$$H^{(2)} = ReLU\left(W^{(2)} \cdot A \cdot H^{(1)}\right)$$
$$\vdots$$
$$H^{(K)} = ReLU\left(W^{(K)} \cdot A \cdot H^{(K-1)}\right)$$
$$x^{(NN)} = \sum_{atoms} H^{(K)}$$
$$h^{(1)} = ReLU\left(W^{(1)} \cdot x^{(NN)}\right)$$
$$h^{(2)} = ReLU\left(W^{(2)} \cdot h^{(1)}\right)$$
$$\vdots$$
$$h^{(K)} = ReLU\left(W^{(K)} \cdot h^{(K-1)}\right)$$

Figure 6. Comparison of algorithms for multilayer perceptron versus graph convolution.

While we are restricted with respect to the compounds in our training data that we can disclose, we can share select publicly disclosed compounds that happened to have been tested in the assays discussed in this work. Table S12 lists commercially available compounds for which MT-PotentialNet's predictions are most improved compared to RF's predictions in the temporal split setting. For example, the first compound, methyl 4-chloro-2-iodobenzoate, has an experimental log D of 3.88, RF predicts log D to be 2.26, and MT-PotentialNet predicts log D to be 3.70. Neural network interpretation remains a discipline in its infancy and therefore renders it challenging to pinpoint exactly which aspect of either the initial featurization or the network enables MT-PotentialNet to properly estimate the log D, while RF significantly underestimates it (for preliminary work on feature interpretation, we point the reader to Supporting Information

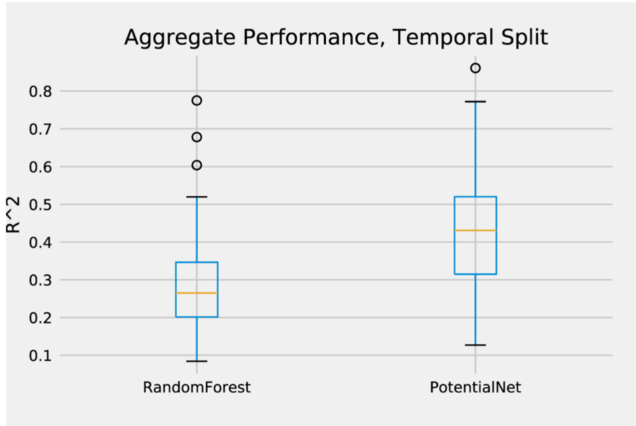


Figure 7. Aggregate performance: temporal split.

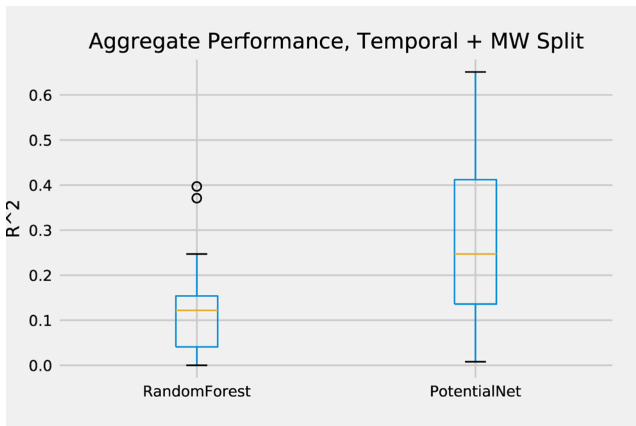


Figure 8. Aggregate performance: Temporal plus MW split.

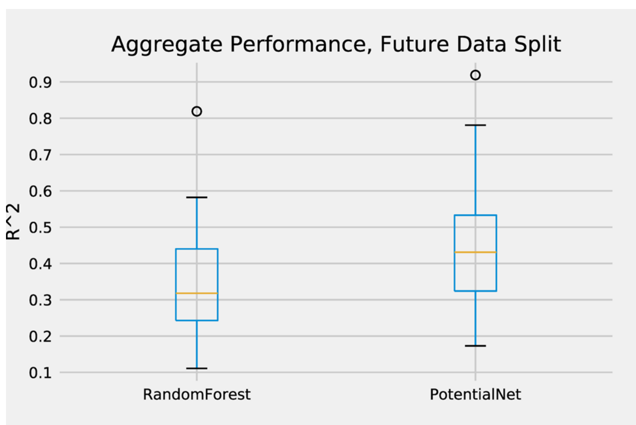


Figure 9. Aggregate performance: evaluation on future data.

https://dx.doi.org/10.1021/acs.jmedchem.9b02187
J. Med. Chem. 2020, 63, 8835−8848

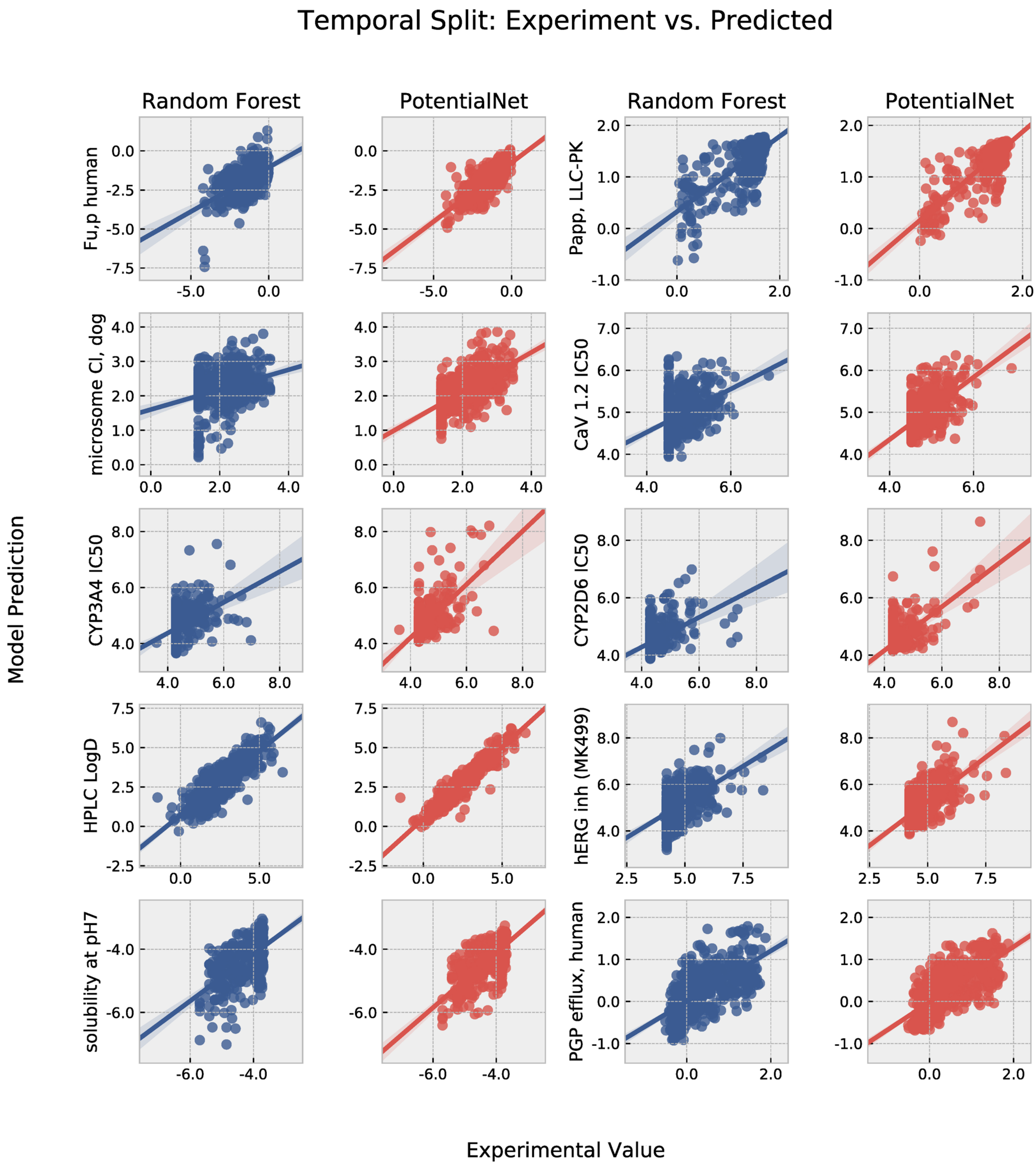


Figure 10. Temporal split: scatter plots of predictions by PotentialNet and by RF vs experiment for several assays.

and also refer to ref 24). As a hypothetical analysis, pair features would include such terms as "carbonyl oxygen that is three bonds away from ether carbon", "carbonyl oxygen that is four bonds away from aromatic iodine", and "carbonyl oxygen that is six bonds away from an aromatic chlorine". There is no sense that it is the *same* carbonyl carbon that has all of these properties. In stark contrast, by recursively propagating information, a graph convolution would confer a single, dense "atom type" on the carbonyl oxygen that would reflect its identity as a halogenated benzaldehyde oxygen.

The differential performance of neural network algorithms versus traditional machine learning methodologies is often attributed to better ability to leverage large data sets. However, a potential corollary of that trend is that such a performance differential may evaporate for smaller data sets. To investigate the effect of data set size on differential performance, we performed a "data ablation" study, in which we progressively removed temporally newer data from the training set while keeping the test set constant. We observed that removing data

https://dx.doi.org/10.1021/acs.jmedchem.9b02187
J. Med. Chem. 2020, 63, 8835−8848

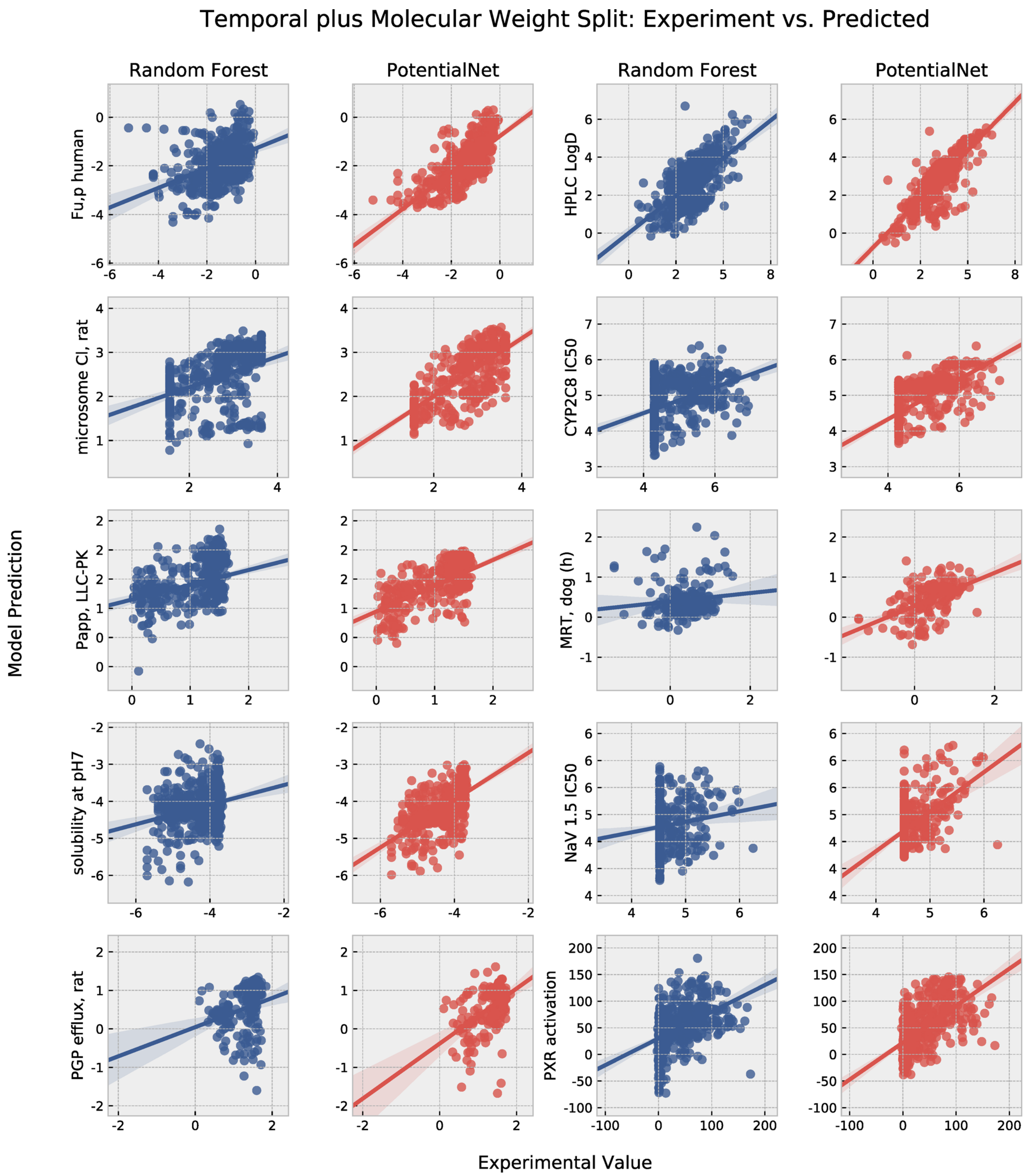


Figure 11. Temporal plus molecular weight split: scatter plots of predictions by PotentialNet and by RF vs experiment for several assays.

from the training set impacted the performance to similar degrees for both random forest and PotentialNet (Figure S5).

More accurate prediction of ADMET end points can be a torchlight[25] guiding creative medicinal chemists as they explore uncharted chemical space en route to the optimal molecule. The results delineated in this paper demonstrate that deep feature learning with graph convolutions and the use of multitask learning in featurization can quite significantly outperform RF based on fixed fingerprints.

■ METHODS

PotentialNet (eq 1)[17] neural networks were constructed and trained with PyTorch.[25] Following previous works,[26] we make extensive use of multitask learning to train our PotentialNet models. We modified the standard multitask framework to save

https://dx.doi.org/10.1021/acs.jmedchem.9b02187
J. Med. Chem. 2020, 63, 8835−8848

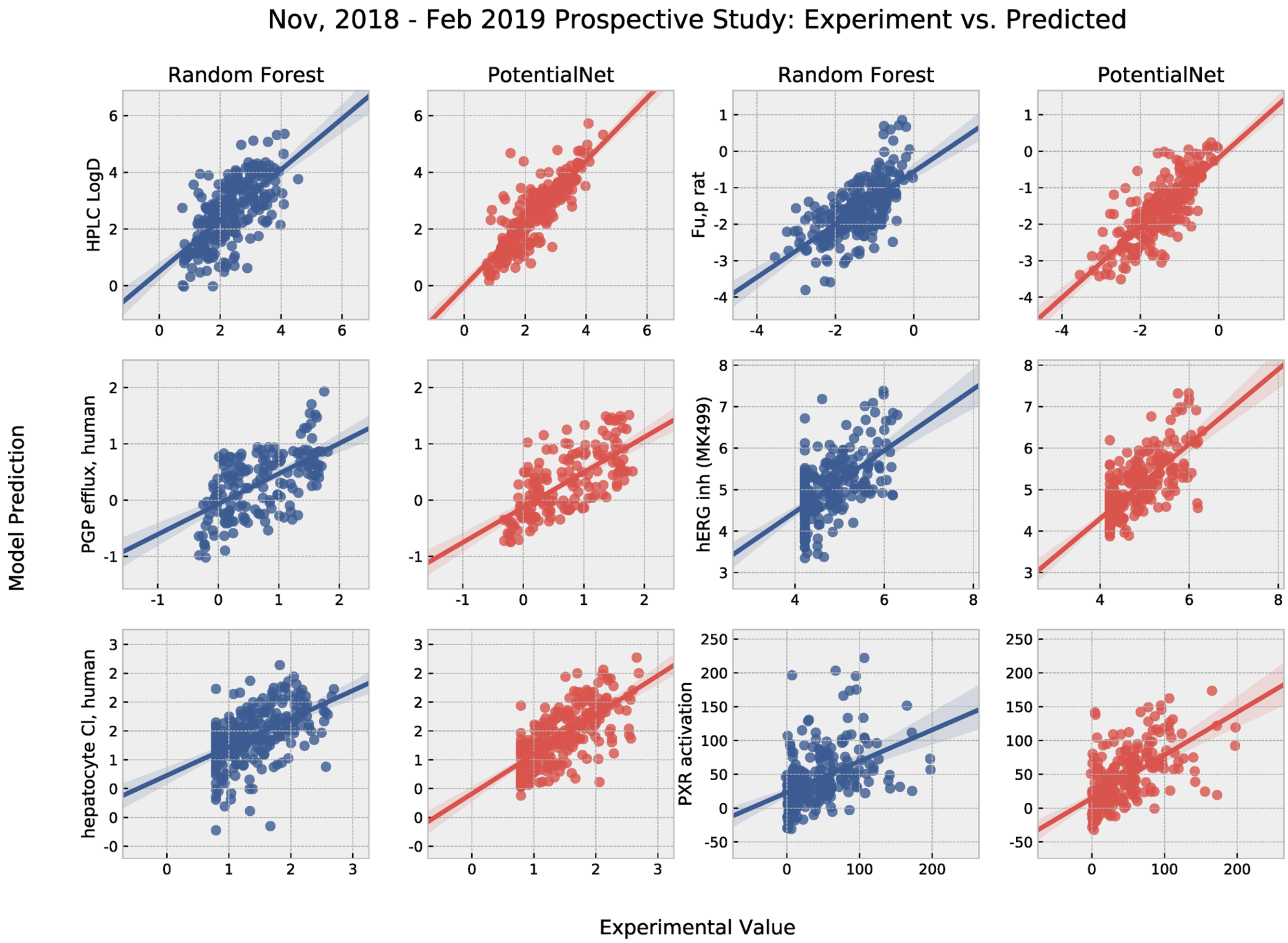


Figure 12. Nov 2018 to Feb 2019 prospective data: scatter plots of predictions by PotentialNet and by RF vs experiment for several assays. All models were trained on data and compounds registered up through Aug 2018 and tested prospectively on data and compounds registered in Nov 2018 to Feb 2019.

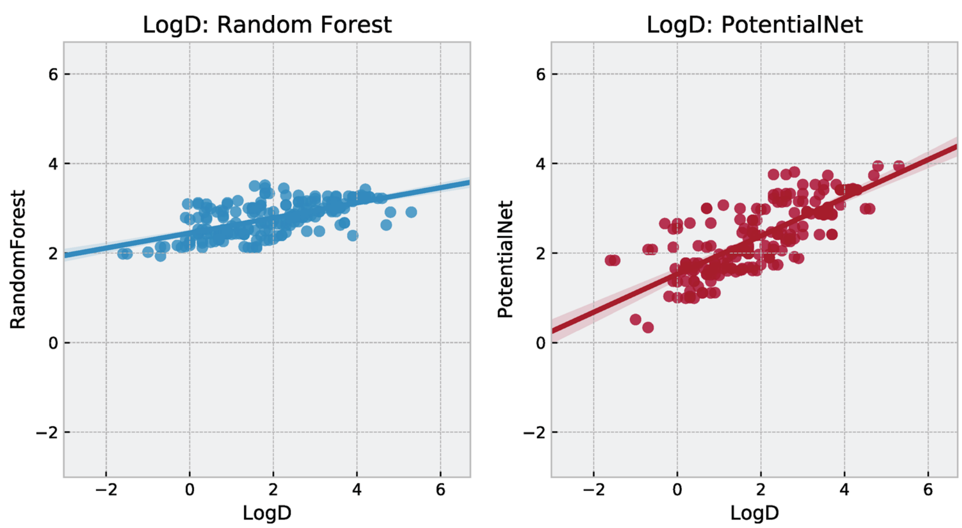


Figure 13. Scatter plots of predictions by models pretrained on Merck data with both PotentialNet and RF vs experiment for log *D* measurements on macrocycles in ref 21.

different models for each task on the epoch at which performance was best for that specific task on the validation set (Figure 1). In that way, we employ an approach that draws on elements of both single and multitask learning. Custom Python code was used based on RDKit[27] and OEChem[28] with frequent use of NumPy[29] and SciPy.[30] Networks were trained on chemical element, formal charge, hybridization, aromaticity, and the total numbers of bonds, hydrogens (total and implicit),

https://dx.doi.org/10.1021/acs.jmedchem.9b02187
J. Med. Chem. 2020, 63, 8835−8848

and radical electrons. RF was implemented using both scikit-learn[31] and MIX; all sklearn-trained random forests models were trained with 500 trees and $\sqrt{n_{\text{features}}}$ per tree; xgboost models were trained using MIX.

To both compare and contrast chemical ML on fixed vector descriptors with chemical deep learning on graph features, we write out a multilayer perceptron (MLP) and a graph convolutional neural network (GCNN) side-by-side (Figure 6). In Figure 6, each molecule is represented either by a flat vector x for the MLP or by both an $N_{\text{atoms}} \times N_{\text{atoms}}$ adjacency matrix A and an $N_{\text{atoms}} \times f_{\text{in}}$ per-atom feature matrix X for the GCNN. The GCNN begins with K graph convolutional layers. It then proceeds to a graph gather operation that sums over the per-atom features in the last graph convolutional hidden layer: $x^{(NN)} = \sum_{\text{atoms}} H^{(K)}$, where the differentiable $x^{(NN)}$, by analogy to the fixed input x of the MLP, is a flat vector graph convolutional fingerprint for the entire molecule, and $H_i^{(K)}$ is the feature map at the K'th graph convolutional layer for atom i. The final layers of the GCNN are identical in form to the hidden layers of the MLP. The difference, therefore, between the MLP and the GCNN lies in the fact that x for MLP is a fixed vector of molecular fingerprints, whereas $x^{(NN)}$ of the GCNN is an end-to-end differentiable fingerprint vector: the features are *learned* in the graph convolution layers.

Another noteworthy parallel arises between the MLP hidden layers and the GCNN graph convolutional layers. Whereas the first MLP layer maps $h^{(1)} = ReLU(W^{(1)} \cdot x)$, the GCNN inserts the adjacency matrix A between W and atom feature matrix X: $H^{(1)} = ReLU(W^{(1)} \cdot A \cdot X)$. Note that while the feature maps X, $H^{(1)}$, ..., $H^{(K)}$ change at each layer of a GCNN, the adjacency matrix A is a constant to be reused at each layer. Therefore, in a recursive manner, a given atom is passed information about other atoms successively further in bond path length at each graph convolutional layer.

Since the advent of the basic graph convolutional neural network, a spate of new approaches[13−17] have improved upon the elementary graph convolutional layers expressed in Figure 6. Here, we train neural networks based on the PotentialNet[17] family of graph convolutions.

$$
\begin{aligned}
h_i^{(1)} &= GRU\left(x_i, \sum_{e}^{N_{\text{et}}} \sum_{j \in N^{(e)}(v_i)} NN^{(e)}(x_j)\right)\\
&\vdots\\
h_i^{(K)} &= GRU\left(h_i^{(b_{K-1})}, \sum_{e}^{N_{\text{et}}} \sum_{j \in N^{(e)}(v_i)} NN^{(e)}(h_j^{(b_{K-1})})\right)\\
h^{(NN)} &= \sigma(i(h^{(K)}, x)) \odot (j(h^{(K)}))\\
h^{(FC_0)} &= \sum_{j=1}^{N_{\text{Lig}}} h_j^{(NN)}\\
h^{(FC_1)} &= ReLU(W^{(FC_1)} h^{(FC_0)})\\
&\vdots\\
h^{(FC_K)} &= W^{(FC_K)} h^{(FC_{K-1})}
\end{aligned}
\tag{1}
$$

where $h_i^{(k)}$ represents the feature map for $atom_i$ at graph convolutional layer k; i, j, and NN are neural networks, N_{Lig} is the number of ligand atoms, and $\{W\}$ are weight matrices for different layers. The GRU is a gated recurrent unit which affords a more efficient passing of information to an atom from its neighbors.

Prior to model training and evaluation, duplicate compounds were removed from the data set, with duplicates uncovered based on OEChem canonical SMILES string equality. To facilitate model training, we standardized training sets to zero mean and unit standard deviation and then subtracted the training set's mean from all values of the test set and divided test set values by the standard deviation of the training set.

QSAR Descriptors: Chemical descriptors, termed "APDP" used in this study for random forests, xgboost, and MLP DNNs, are listed as follows. All descriptors are used in frequency form; i.e., we use the number of occurrences in a molecule and not just the binary presence or absence. APDP denotes the union of AP, the original "atom pair" descriptor from ref 9, and DP descriptors ("donor−acceptor pair"), called "BP" in ref 32. Such APDP descriptors are used in most of Merck's QSAR studies and in Merck's production QSAR. Both descriptors are of the form "atom $type_i$-(distance in bonds)-atom $type_j$".

For AP, atom type includes the element, number of nonhydrogen neighbors, and number of π electrons; it is very specific. For DP, atom type is one of seven (cation, anion, neutral donor, neutral acceptor, polar, hydrophobe, and other); it contains a more generic description of chemistry.

QSAR Methods: All methods are used in regression mode; i.e., both input activities and predictions are floating-point numbers. All appropriate descriptors are used in the models; i.e., no feature selection is done. When random forests are not trained with scikit-learn, they are trained with the Merck MIX library that in turn calls the R module RandomForest,[33] which encodes the original method of ref 34 and first applied to QSAR in ref 35. The default settings are 100 trees, nodesize = 5, mtry = $M/3$ where M is the number of unique descriptors.

MLP Deep Neural Networks (DNN): We use Python-based code obtained from the Kaggle contest and described in ref 36. In this paper, cf. Supporting Information Table S5, we use parameters slightly different from the "standard set" described in that paper: two intermediate layers of 1000 and 500 neurons with 25% dropout rate and 75 training epochs. The above change is made for the purposes of more time-efficient calculation. The accuracy of prediction is very similar to that of the standard set.

xgboost: Extreme gradient boosting method is published in ref 37. In this paper, cf. Supporting Information Table S5, we are using a set of standard parameters from Merck's subsequent study using this method.[6,38]

ADME Assays: The experimental approaches for the assays shown here are detailed in previous publications from Merck & Co.[5]

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jmedchem.9b02187.

> A feature interpretation approach; additional temporal split and temporal plus molecular weight split analysis figures and tables (PDF)

https://dx.doi.org/10.1021/acs.jmedchem.9b02187
J. Med. Chem. 2020, 63, 8835−8848

AUTHOR INFORMATION

Corresponding Authors

Evan N. Feinberg – *Program in Biophysics, Stanford University, Palo Alto, California 94305, United States; Computational and Structural Chemistry, Merck & Co., Inc., South San Francisco, California 94080, United States;* orcid.org/0000-0002-7989-8751; Email: evan.n.feinberg@gmail.com

Alan C. Cheng – *Computational and Structural Chemistry, Merck & Co., Inc., South San Francisco, California 94080, United States;* orcid.org/0000-0003-3645-172X; Email: alan.cheng@merck.com

Authors

Elizabeth Joshi – *Pharmacokinetics, Pharmacodynamics, and Drug Metabolism, Merck & Co., Inc., Kenilworth, New Jersey 07065, United States;* orcid.org/0000-0003-3267-0634

Vijay S. Pande – *Department of Bioengineering, Stanford University, Palo Alto, California 94305, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jmedchem.9b02187

Notes

The authors declare the following competing financial interest(s): The authors are currently employed by Merck & Co. and Genesis Therapeutics, Inc.

ACKNOWLEDGMENTS

We thank Juan Alvarez, Scott Johnson, Andy Liaw, Robert Sheridan, Matthew Tudor, Isha Verma, and Yuting Xu for their helpful comments and insightful discussions in preparing this manuscript. We are grateful to the anonymous reviewers for their suggestions. We acknowledge the support of the Blue Waters Graduate Fellowship. The Pande Group acknowledges the generous support of Dr. Anders G. Frøseth and Christian Sundt for our work on machine learning. And finally, we are grateful for the support of Merck & Co. for supporting this work.

ABBREVIATIONS USED

ML, machine learning; RF, random forest; MLP, multilayer perceptron; GCNN, graph convolutional neural network; DNN, deep neural network; EPSA, experimental polar surface area

REFERENCES

(1) Kennedy, T. Managing the Drug Discovery/Development Interface. *Drug Discovery Today* **1997**, *2*, 436–444.

(2) Kola, I.; Landis, J. Can the Pharmaceutical Industry Reduce Attrition Rates? *Nat. Rev. Drug Discovery* **2004**, *3*, 711.

(3) Loving, K. A.; Lin, A.; Cheng, A. C. Structure-Based Druggability Assessment of the Mammalian Structural Proteome With Inclusion of Light Protein Flexibility. *PLoS Comput. Biol.* **2014**, *10*, e1003741.

(4) Cheng, A. C.; Eksterowicz, J.; Geuns-Meyer, S.; Sun, Y. Analysis of Kinase Inhibitor Selectivity Using a Thermodynamics-Based Partition Index. *J. Med. Chem.* **2010**, *53*, 4502–4510.

(5) Sherer, E. C.; Verras, A.; Madeira, M.; Hagmann, W. K.; Sheridan, R. P.; Roberts, D.; Bleasby, K.; Cornell, W. D. QSAR Prediction of Passive Permeability in the Llc-Pk1 Cell Line: Trends in Molecular Properties and Cross-Prediction of Caco-2 Permeabilities. *Mol. Inf.* **2012**, *31*, 231–245.

(6) Sanders, J. M.; Beshore, D. C.; Culberson, J. C.; Fells, J. I.; Imbriglio, J. E.; Gunaydin, H.; Haidle, A. M.; Labroli, M.; Mattioni, B. E.; Sciammetta, N.; Shipe, W. D.; Sheridan, R. P.; Suen, L. M.; Verras, A.; Walji, A.; Joshi, E. M.; Bueters, T. Informing the Selection of Screening Hit Series with in Silico Absorption, Distribution, Metabolism, Excretion, And Toxicity Profiles. *J. Med. Chem.* **2017**, *60*, 6771–6780.

(7) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer: New York, NY, U.S., 2009; pp 9–41.

(8) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–754.

(9) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure-Activity Studies: Definition and Applications. *J. Chem. Inf. Model.* **1985**, *25*, 64–73.

(10) Goodfellow, I.; Bengio, Y.; Courville, A.; Bengio, Y. *Deep Learning*; MIT Press: Cambridge, MA, 2016; Vol. *1*.

(11) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of Mdl Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1273–1280.

(12) Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. In *Advances in Neural Information Processing Systems 28*, Montréal, Canada, December 7–12, 2015; Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NY, 2015; pp 2224–2232.

(13) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular Graph Convolutions: Moving Beyond Fingerprints. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 595–608.

(14) Kipf, T. N.; Welling, M. Semi-Supervised Classification With Graph Convolutional Networks. *arXiv* **2016**, arXiv:1609.02907.

(15) Li, Y.; Zemel, R.; Brockschmidt, M.; Tarlow, D. *Proceedings*, International Conference on Learning Representations, San Juan, Puerto Rico, May 2–4, 2016; ICLR, 2016.

(16) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. *Proceedings*, 34th International Conference on Machine Learning, International Convention Centre, Sydney, Australia, 2017; International Machine Learning Society, 2017; pp 1263–1272.

(17) Feinberg, E. N.; Sur, D.; Wu, Z.; Husic, B. E.; Mai, H.; Li, Y.; Sun, S.; Yang, J.; Ramsundar, B.; Pande, V. S. PotentialNet for Molecular Property Prediction. *ACS Cent. Sci.* **2018**, *4*, 1520–1530.

(18) Sheridan, R. P. Time-Split Cross-Validation as a Method for Estimating the Goodness of Prospective Prediction. *J. Chem. Inf. Model.* **2013**, *53*, 783–790.

(19) Walters, P. Solubility. https://github.com/PatWalters/solubility, 2018.

(20) Xu, Y.; Ma, J.; Liaw, A.; Sheridan, R. P.; Svetnik, V. Demystifying Multitask Deep Neural Networks for Quantitative Structure–Activity Relationships. *J. Chem. Inf. Model.* **2017**, *57*, 2490–2504.

(21) Over, B.; Matsson, P.; Tyrchan, C.; Artursson, P.; Doak, B. C.; Foley, M. A.; Hilgendorf, C.; Johnston, S. E.; Lee, M. D.; Lewis, R. J.; McCarren, P.; Muncipinto, G.; Norinder, U.; Perry, M. W. D.; Duvall, J. R.; Kihlberg, J. Structural and Conformational Determinants of Macrocycle Cell Permeability. *Nat. Chem. Biol.* **2016**, *12*, 1065–1074.

(22) Wager, T. T.; Chandrasekaran, R. Y.; Hou, X.; Troutman, M. D.; Verhoest, P. R.; Villalobos, A.; Will, Y. Defining Desirable Central Nervous System Drug Space Through the Alignment of Molecular Properties, in Vitro Adme, and Safety Attributes. *ACS Chem. Neurosci.* **2010**, *1*, 420–434.

(23) Segall, M.; Champness, E.; Leeding, C.; Lilien, R.; Mettu, R.; Stevens, B. Applying Medicinal Chemistry Transformations and Multiparameter Optimization to Guide the Search for High-Quality Leads and Candidates. *J. Chem. Inf. Model.* **2011**, *51*, 2967–2976.

(24) McCloskey, K.; Taly, A.; Monti, F.; Brenner, M. P.; Colwell, L. J. Using Attribution to Decode Binding Mechanism in Neural Network Models for Chemistry. *Proc. Natl. Acad. Sci. U. S. A.* **2019**, *116*, 11624–11629.

(25) Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; Lerer, A. Automatic differentiation in PyTorch. In *Neural Information Processing Systems Autodiff Workshop*, Long Beach, CA, U.S., December 9, 2017; Wiltschko, A., van Merriënboer, B., Lamblin, P., Eds.; Autodiff, 2017;

https://openreview.net/pdf?id=BJJsrmfCZ (accessed September 10, 2018).
(26) Ramsundar, B.; Kearnes, S.; Riley, P.; Webster, D.; Konerding, D.; Pande, V. Massively Multitask Networks for Drug Discovery. *arXiv* **2015**, arXiv:1502.02072.
(27) RDKit: Open-source cheminformatics. http://www.rdkit.org (accessed September 10, 2018).
(28) *OEChem*; OpenEye Scientific Software: Santa Fe, NM; http://www.eyesopen.com/ (accessed September 10, 2018).
(29) van der Walt, S.; Colbert, S. C.; Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. *Comput. Sci. Eng.* **2011**, *13*, 22−30.
(30) Jones, E.; Oliphant, T.; Peterson, P. SciPy: Open Source Scientific Tools for Python, 2001−. http://www.scipy.org/ (accessed September 10, 2018).
(31) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(32) Kearsley, S. K.; Sallamack, S.; Fluder, E. M.; Andose, J. D.; Mosley, R. T.; Sheridan, R. P. Chemical Similarity Using Physiochemical Property Descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 118−127.
(33) Breiman, L.; Cutler, A.; Liaw, A.; Wiener, M. Package Randomforest. https://www.stat.berkeley.edu/~breiman/RandomForests/.
(34) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5−32.
(35) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and Qsar Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947−1958.
(36) Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V. Deep Neural Nets as a Method for Quantitative Structure−Activity Relationships. *J. Chem. Inf. Model.* **2015**, *55*, 263−274.
(37) Chen, T.; Guestrin, C. Xgboost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; Association for Computing Machinery: New York, 2016; pp 785−794.
(38) Sheridan, R. P.; Wang, W. M.; Liaw, A.; Ma, J.; Gifford, E. M. Extreme Gradient Boosting as a Method for Quantitative Structure−Activity Relationships. *J. Chem. Inf. Model.* **2016**, *56*, 2353−2360.

https://dx.doi.org/10.1021/acs.jmedchem.9b02187
J. Med. Chem. 2020, 63, 8835−8848