JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Modeling a Crowdsourced Definition of Molecular Complexity

Robert P. Sheridan,*[,†] Nicolas Zorn,[‡] Edward C. Sherer,[†] Louis-Charles Campeau,[§] Charlie (Zhenyu) Chang,[∥] Jared Cumming,[⊥] Matthew L. Maddess,[#] Philippe G. Nantermet,[▲] Christopher J. Sinz,[●] and Paul D. O'Shea[○]

[†]Structural Chemistry, Merck Research Laboratories, Merck & Co., Inc., P.O. Box 2000, Rahway, New Jersey 07065, United States

[‡]Structural Chemistry, Merck Research Laboratories, Merck & Co., Inc., 2000 Galloping Hill Road, Kenilworth, New Jersey 07033, United States

[§]Process Chemistry, Merck Research Laboratories, Merck & Co., Inc., P.O. Box 2000, Rahway, New Jersey 07065, United States

[∥]Structural Chemistry, Merck Research Laboratories, Merck & Co., Inc., 33 Avenue Louis Pasteur, Boston, Massachusetts 02115, United States

[⊥]Discovery Chemistry, Merck Research Laboratories, Merck & Co., Inc., 2000 Galloping Hill Road, Kenilworth, New Jersey 07033, United States

[#]Process Chemistry, Merck Research Laboratories, Merck & Co., Inc., 33 Avenue Louis Pasteur, Boston, Massachusetts 02115, United States

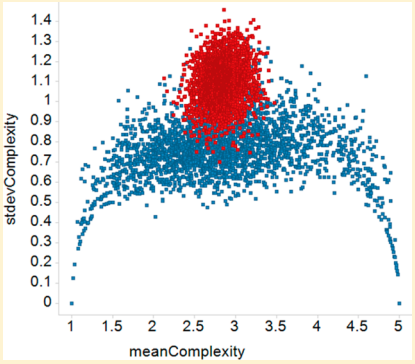
[▲]Discovery Chemistry, Merck Research Laboratories, Merck & Co., Inc., 770 Sumneytown Pike, West Point, Pennsylvania 19486, United States

[●]Discovery Chemistry, Merck Research Laboratories, Merck & Co., Inc., P.O. Box 2000, Rahway, New Jersey 07065, United States

[○]Analytical Chemistry, Merck Research Laboratories, Merck & Co., Inc., P.O. Box 2000, Rahway, New Jersey 07065, United States

Supporting Information

ABSTRACT: This paper brings together the concepts of molecular complexity and crowdsourcing. An exercise was done at Merck where 386 chemists voted on the molecular complexity (on a scale of 1−5) of 2681 molecules taken from various sources: public, licensed, and in-house. The meanComplexity of a molecule is the average over all votes for that molecule. As long as enough votes are cast per molecule, we find meanComplexity is quite easy to model with QSAR methods using only a handful of physical descriptors (e.g., number of chiral centers, number of unique topological torsions, a Wiener index, etc.). The high level of self-consistency of the model (cross-validated R^2 ~0.88) is remarkable given that our chemists do not agree with each other strongly about the complexity of any given molecule. Thus, the power of crowdsourcing is clearly demonstrated in this case. The meanComplexity appears to be correlated with at least one metric of synthetic complexity from the literature derived in a different way and is correlated with values of process mass intensity (PMI) from the literature and from in-house studies. Complexity can be used to differentiate between in-house programs and to follow a program over time.



INTRODUCTION

Chemists have an intuitive idea whether a molecule is "simple" or "complex". A number of proposals have been made in the literature[1−15] to define this more objectively. However, different definitions often emphasize different aspects of the molecule. As with any other type of intuitive concept, any given definition of complexity can seem reasonable to the person who proposed it, but completely arbitrary to others. Authors vary on how much the concept of "complexity" is a surrogate for "synthetic difficulty". A priori, we would expect that these are separate concepts, although they may correlate. For the purposes of this paper, we think of complexity of a molecule as a property defined totally by its chemical structure. Synthetic difficulty, on the other hand, will depend on what agents are available and what synthetic routes are available, something that varies from laboratory to laboratory and with time.[6]

Merck process chemists assign "process complexity" to a molecule as a way of gauging the effort required for either small- or large-scale synthesis. One other desired use of complexity is to help measure the relative efficiency of two synthetic efforts when no molecules are in common, for instance when comparing different companies. One may measure the overall effort of the synthesis of a given molecule by the number of steps, process mass intensity, etc., but to compare two different molecules one must normalize the effort by the expected difficulty for each

Received: March 20, 2014

   dx.doi.org/10.1021/ci5001778 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

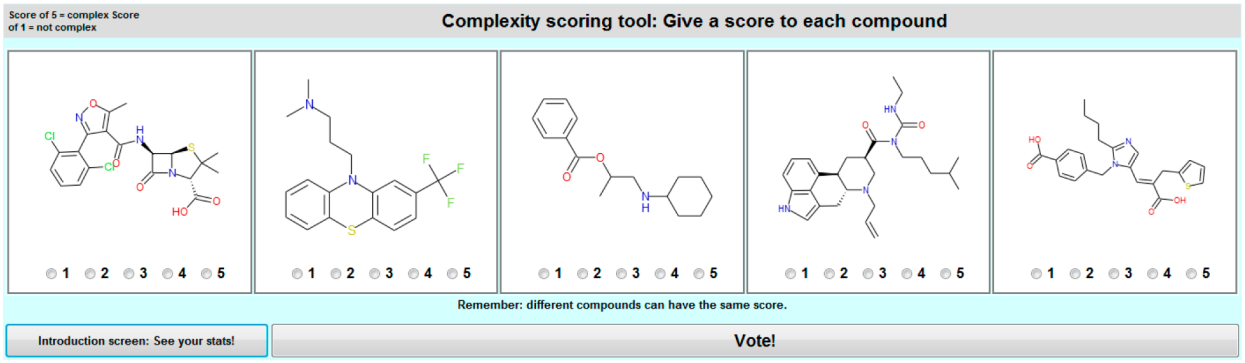


Figure 1. Example voting window.

molecule. Complexity could be taken as a surrogate of "the expected difficulty" for that purpose.

Our study grew from the realization that, rather than depending on a few chemists to assign complexity for a few molecules "by hand", it would be desirable to have an objective way to assign complexity to a large number of molecules whatever the source. It was decided that assignment of complexity should be done with a single QSAR model and the model should be calibrated against a large set of molecules. The assignment of complexity for each molecule should, in turn, represent the consensus over a large number of chemists throughout Merck. Getting the consensus would thus require an in-house "crowdsourcing" exercise. Implicit to this type of crowdsourcing[12,16−18] is the idea that, although different chemists might disagree about an ill-defined or subjective property of a molecule, the average over many chemists would approach something consistent and useful.

This paper describes the crowdsourcing exercise to define complexity within Merck and the method of generating a model of complexity. We show some preliminary indications that our complexity definition correlates with at least one synthetic accessibility definition in the literature and correlates with process mass index from the literature and from in-house efforts.

SOURCE OF MOLECULES

The complexity training set contained molecules from varied sources, some public, some licensed, and some in-house.

1. 1192 randomly selected molecules from the MDDR.[19]
2. 77 randomly selected from ChEMBL.[20]
3. 200 "top drugs" subset.[21]
4. 40 molecules from Ertl and Schuffenhauer[9] a previously published paper on complexity.
5. 14 molecules from Kjell et al.[15]
6. 272 "natural products from the ZINC database.[22]
7. 148 late stage Merck molecules.
8. 554 molecules synthesized by the Merck Discovery Process Chemistry group (DPC molecules) in support of SAR activities.
9. 154 randomly selected Merck molecules.
10. Two sets of randomly selected molecules from sets 1−9 that were duplicated to test voters' perception of complexity due to rotation of the drawing or indications of chirality.

The total number of molecules is 2681, of which 2575 are unique. As indicated, some duplication was added deliberately in set 10. Other duplications are due to overlapping compound sets, e.g. some molecules in "top drugs" were also in complexity data sets from the literature. Duplicates were treated as separate molecules for the purposes of voting.

THE VOTING PHASE

A voting tool was created with Accelrys Pipeline Pilot 8.5[23] and deployed in-house as a Web application. It was designed to maximize user engagement and voting efficiency and to provide robust data collection and monitoring. An initial landing page was designed to provide voters with the goals and details on the voting process as well as show the latest voting statistics. Explicit instructions stated that the goal was to score "complexity" and not specifically synthetic difficulty. Chemists could then open the voting page and assign complexity scores on a scale of 1 to 5 to each of a set of five randomly selected molecules. A score of 1 would be "least complex", and a score of 5 would be "most complex". An example of a voting page is shown in Figure 1.

Once the votes were recorded, another set of five randomly selected molecules appeared. This continued until the chemist ended the session. Each ballot (i.e., one set of five molecules scored on a page) was stored in an Oracle database where the following data was captured: molecule ID and score, voter ID, ballot date and time, ballot voting duration, and ballot number in the voting session. To ensure that all molecules were scored (and prevent unproductive oversampling of any given molecule), molecules voted on more than a fixed number of times were removed at the end of each day from the pool of molecules to be scored.

The voting tool was set up to send email reminders to individual voters and included reports of the progress of each individual and the cumulative progress of all the voters.

We recruited chemists at four Merck sites representing a variety of subdisciplines (process chemists, analytical chemists, medicinal chemists, computational chemists, crystallographers, etc.). Chemists were encouraged to cast at least 250 votes each. Overall ∼108,000 votes from 386 chemists for 2681 molecules were compiled over a period of 38 days. This comes to an average of ∼300 votes per chemist, more than expected. The highest number of votes cast by an individual chemist was 1591. Given the stochastic way the molecules were presented, and given the number of chemists, molecules, and mean number of votes per chemist, different chemists saw a different subset of molecules. A molecule might be voted on once or several times by the same chemist. On average a molecule received 41 ± 16 votes. The assumption, which will be later seen to be mostly correct, is that getting a self-consistent model for complexity depends strongly on the average number of votes per molecule. Moreover, it is not necessary for the same set of chemists to score every molecule.

dx.doi.org/10.1021/ci5001778 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

MEANCOMPLEXITY AND STDEVCOMPLEXITY COMPARED TO THE NULL HYPOTHESIS

We will use meanComplexity of a molecule to indicate the mean value of all votes for that molecule, and stdevComplexity to mean the standard deviation of all votes for that molecule. The meanComplexity over all molecules is roughly normally distributed with a mean of 2.85 ± 0.78. The mean stdevComplexity over all molecules is 0.77 ± 0.14.

One early concern was that "complexity" was such a subjective concept that chemists would disagree with each other (and with themselves on separate occasions of voting on the same molecule) to such an extent that it would be impossible to distinguish less complex from more complex molecules based on consensus. We therefore created, for the purpose of comparison, several "null hypothesis" data sets where individual vote values (1 through 5) were randomly reassigned to individual votes. This represents the worst case situation where any agreement was due solely to chance.

Figure 2 represents the relationship of stdevComplexity to meanComplexity in two different ways. The real vote is shown in blue, and one null hypothesis data set is shown in red.

In Figure 2 (top) molecules are ordered by increasing meanComplexity along the *x*-axis and three points are shown per molecule: meanComplexity, meanComplexity+stdevComplexity, and meanComplexity-stdevComplexity. In Figure 2 (bottom) stdevComplexity is plotted vs meanComplexity.

In both figures, the real vote is clearly different from the null hypothesis. The meanComplexity for the null hypothesis never varies far from ∼2.89 and the mean stdevComplexity over all molecules is ∼1.3 instead of 0.77. Thus, voters are distinguishing less complex from more complex molecules much better than in the null hypothesis. However, it is clear that stdevComplexity of the real vote is a large fraction of the range in meanComplexity, indicating that chemists do not closely agree with each other about any given molecule. However, later we will see that it is still possible to make a very self-consistent model of meanComplexity.

The stdevComplexity in the real vote is more or less constant except at the upper and lower range of meanComplexity; it falls toward zero when meanComplexity <1.5 or meanComplexity >4.5. (This is easier to see in Figure 2 bottom.) This is a mathematical consequence of having the lower and upper scores being 1 and 5. For example, to approach a meanComplexity of 1, all the votes for a molecule would have to be very close to 1, i.e. have a small stdevComplexity for that molecule.

THE EFFECT OF ROTATION AND CHIRALITY DEPICTION IN THE PERCEPTION OF COMPLEXITY

The subjective complexity of molecular structure may depend upon several factors. Two factors we chose to explicitly examine are the effect of rotation of the figure and the explicit depiction of chirality. The effect of orientation of the molecule was tested by taking 20 example structures and rotating the image randomly. The comparison of the vote of the initial structure vs the rotated version is shown as the green line in Figure 3. This line is very close to the diagonal (average change 0.06 ± 0.21 for 20 molecules), so rotation did not appear to influence perceived complexity to any significant extent. When 20 molecules that had explicit chirality were edited to remove the hashes and wedges, there was a systematic decrease (average change = −0.35 ± 0.21) in the assigned complexity by chemists (blue line in Figure 3). When hashes and wedges were added for

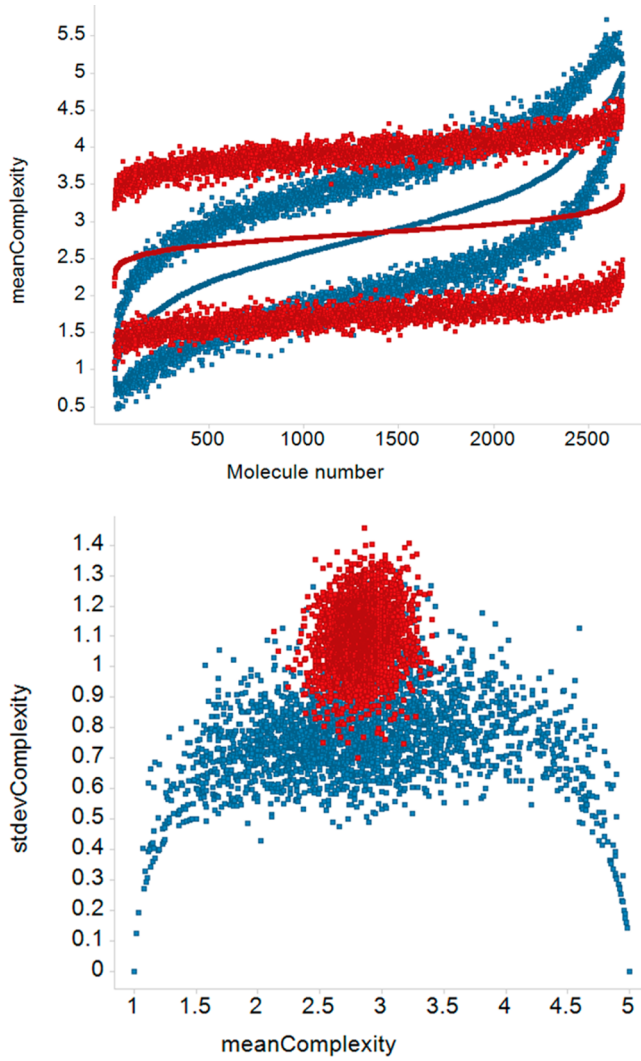


Figure 2. Two different representations of meanComplexity vs stdevComplexity. In each case the blue points represent the results of the voting. The red points represent a "null hypothesis" formed by randomly permuting the correspondence of the vote value (1 through 5) to an individual vote. Top: The molecules are ordered by increasing meanComplexity, and we are showing three points for each molecule: meanComplexity, meanComplexity+stdevComplexity, and meanComplexity-stdevComplexity. Bottom: stdevComplexity vs meanComplexity.

20 molecules initially drawn "racemic", the effect was a systematic increase (average change = 0.44 ± 0.16) in the assigned complexity (red line in Figure 3). Thus, explicit depiction of chirality has a noticeable effect. However, these changes are less the stdevComplexity for an average molecule (0.77). That is, the variation among users would normally overwhelm the effect of the chirality notation. Thus, we are not overly concerned about exactly how a compound is depicted.

HOW MUCH DO INDIVIDUAL CHEMISTS AGREE WITH THE CONSENSUS MEANCOMPLEXITY?

We can examine the how well the votes of individual chemists agree with the consensus meanComplexity. One metric for this is the root-mean-square error over a set of molecules (RMSE). Using RMSE is preferable to using correlations because a correlation can be high (i.e., the trend could be in the same direction), but the numerical values might not match. Note that any given chemist would have scored a small subset of the

dx.doi.org/10.1021/ci5001778 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

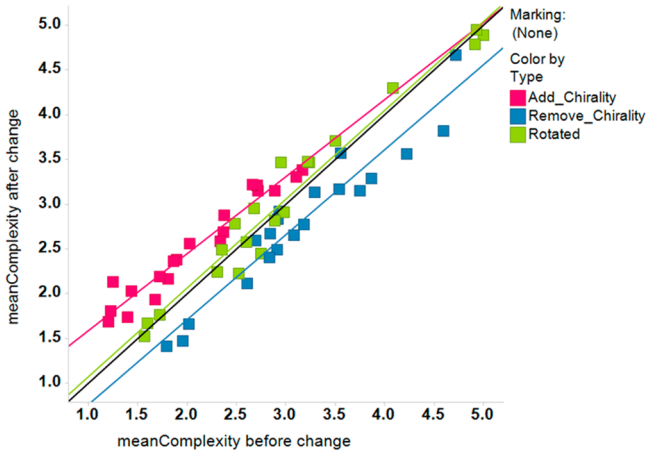


Figure 3. Effect on meanComplexity of rotation of the drawing (green points) and adding (red points) or removing (blue points) wedges and hashes.

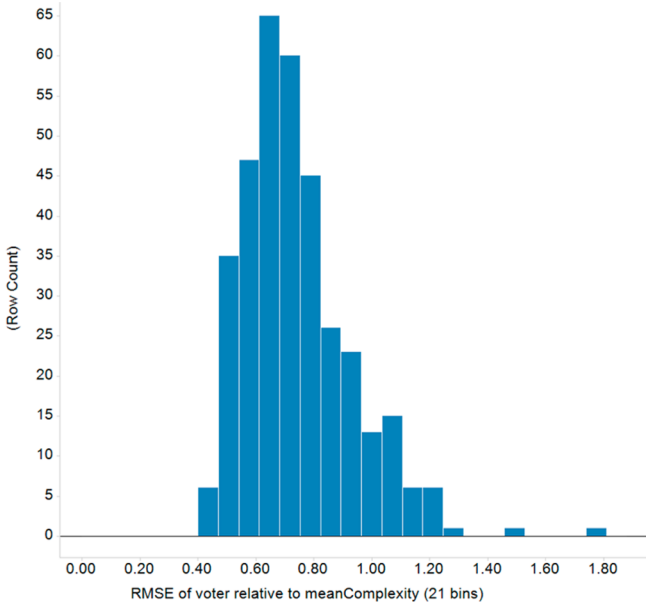


Figure 4. Histogram for the root-mean-square-error between the votes of individual chemists and meanComplexity (the average votes for all chemists), where the voter has scored ≥30 molecules. The mean of this distribution is 0.75, compared to the standard deviation of meanComplexity = 0.77.

molecules (on the average 38 molecules). We will include only those chemists that voted on 30 or more molecules. Figure 4 shows a histogram of the RMSE for all chemists. The mean RMSE is 0.75 compared to the stdev in meanComplexity of 0.77. Some examples of individual votes vs meanComplexity are in Figure 5. Chemist1 has the highest RMSE and voted in a way anticorrelated to the consensus. Later we learned that Chemist1 had misinterpreted the instructions and reversed the complexity scale. Chemist2 systematically voted molecules more complex than the consensus. Chemist3 systematically voted molecules less complex than consensus. Chemist4 had a close to mean RMSE, i.e. was the most typical. Chemist5 had the smallest RMSE, i.e. voted closest to the consensus.

We note that the meanComplexity does not appreciably change whether one includes, excludes, or corrects the votes of Chemist1 (the Pearson correlation over all molecules among these versions of meanComplexity >0.999). This demonstrates the robustness of the consensus against issues with individual chemists.

QSAR MODEL OF MEANCOMPLEXITY

Ultimately we need to generate a model with which we can assign a complexity to any arbitrary molecule (which we will refer to as the "predicted meanComplexity"). The most straightforward way of doing that is to use a QSAR method to fit the meanComplexity against a relevant set of descriptors. For descriptors we tried the following:

AP Carhart atom pairs, which encode detailed atom types and distances.[24]

SP3CARBONS Counts of chiral atoms, sp^3 hybridized atoms, aromatic atoms, and those counts normalized by the total number of non-hydrogen atoms.

DESCRIPTORCOMPLEXITY Ratio of unique descriptors vs total descriptors. For descriptors we used the AP[24] and TT descriptors.[25]

MOE_2D A diverse set of computable physical and topological properties calculated by the software package MOE.[26]

For QSAR methods, we used random forest (RF)[27,28] and linear kernel support vector machine (SVM),[29,30] which are generally held to produce very good QSAR models.[31,32]

HOW SELF-CONSISTENT IS THE MODEL FOR MEANCOMPLEXITY?

The standard way of determining the best descriptor set and determining the self-consistency of a data set is by cross-validation. Half of the molecules are randomly selected as the training set. A QSAR model is built from the training set, and the model is used to predict the test set, which consists of the remaining molecules. The metric R^2 is used to measure the agreement between the observed and predicted meanComplexity of the test set. Here we show the mean R^2 over five such divisions. An R^2 of 0 would mean no self-consistency in the data, and an R^2 of 1.0 would mean perfect consistency.

Cross-validated R^2 ($CV\text{-}R^2$) for several combinations of QSAR method and subsets of descriptors is shown in Table 1. Some descriptor combinations do an excellent job of predicting meanComplexity, with $CV\text{-}R^2 \geq 0.87$. This is much better than we have seen for almost any QSAR data set. The minimum descriptor types needed for maximum $CV\text{-}R^2$ are three: MOE_2D, SP3CARBONS, and DESCRIPTORCOMPLEXITY. The fact that including AP descriptors (or any other type of substructure-based descriptors) makes no improvement in the $CV\text{-}R^2$ over those three descriptor types means, somewhat counterintuitively, that the presence of specific chemical groups do not contribute to the overall impression of complexity. The difference of $CV\text{-}R^2$ between RF and SVM using any given set of descriptor types is probably not significant. Our discussion from now on will refer to the RF model.

A plot of meanComplexity vs the cross-validated prediction of meanComplexity using the RF model and the three descriptors is shown in Figure 6. The root-mean-square deviation for this set of predictions is 0.27, much smaller than the standard deviation of meanComplexity (0.77).

An alternative check of self-consistency is "time-split",[33] i.e. making a model using molecules early in the voting and predicting molecules late in the voting. In this case DPC molecules were exposed for voting after the others, so they will be the test set. The training set will be all other molecules. We will use the three descriptor types we know are minimal in the

dx.doi.org/10.1021/ci5001778 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

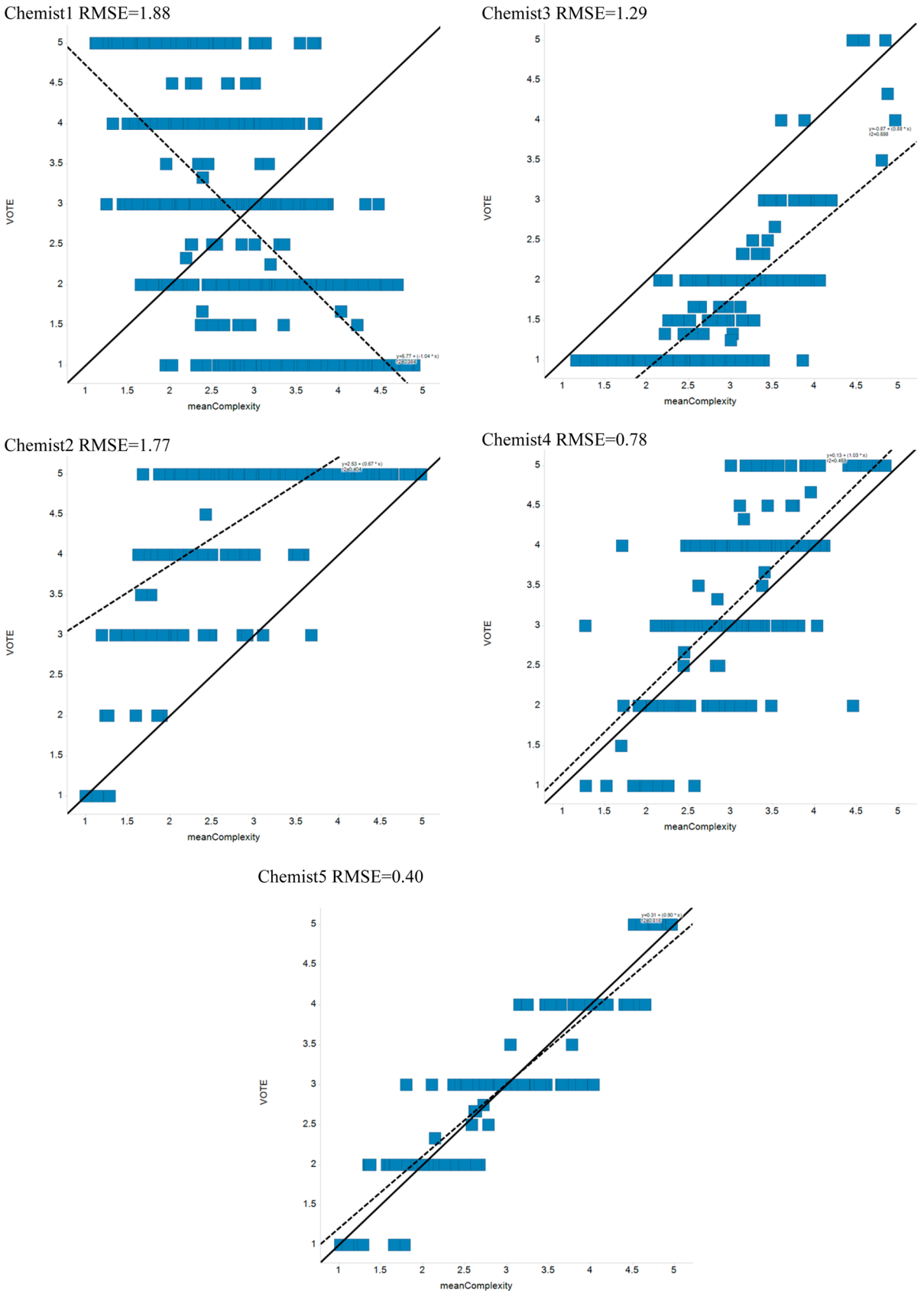


Figure 5. Examples of the votes of individual chemists compared to meanComplexity (i.e., the average vote over all chemists). Each point represents a molecule. The votes may be nonintegers if the chemist voted on the same compound more than once.

random cross-validation discussed in the last paragraph. R^2 for the DPC compounds is shown in Table 1. Again we see $R^2 \geq 0.85$.

The default modeling results in Table 1 include the duplicate molecules because each was voted on as a separate entity.

dx.doi.org/10.1021/ci5001778 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 1. Cross-Validated R^2 and Prospective R^2 for meanComplexity

QSAR method	AP	MOE_2D	SP3CARBONS	DESCRIPTORCOMPLEXITY	CV-R^2
RF	X	X	X	X	**0.88**
RF		X	X	X	**0.88**
RF		X		X	0.85
RF	X				0.80
RF		X			0.85
RF			X		0.77
RF				X	0.64
SVM	X	X	X	X	**0.87**
SVM		X	X	X	**0.87**
SVM		X		X	0.83
SVM		X			0.82
SVM			X		0.74
SVM				X	0.52
					time-split R^2
RF		X	X	X	0.85
SVM		X	X	X	0.86

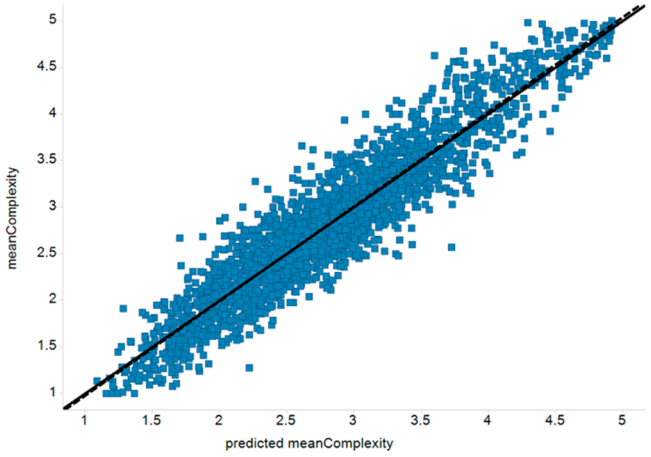


Figure 6. meanComplexity from the voting vs the cross-validated prediction of meanComplexity using the random forest model with MOE_2D, SP3CARBONS, and DESCRIPTORCOMPLEXITY as descriptors. Each point represents the mean of 5 cross-validated predictions. The solid black line represents the diagonal, and the dashed line represents the best least-squares fit.

The CV-R^2 and time-split R^2 using only the 2575 unique molecules in our set are effectively identical (higher by 0.00 to 0.01) to those shown in Table 1.

DEPENDENCY OF CV-R^2 ON VOTES PER MOLECULE

CV-R^2 strongly depends on the average number of votes per molecule. We generated randomly selected subsets of votes of various sizes and for each subset calculated meanComplexity for each molecule. The CV-R^2 as a function of the ratio of votes per molecule for these subsets is shown in Figure 7 (top). Clearly a minimum ratio is needed (>30) to reach the CV-R^2 we see in the full set. We infer this is because we need that many votes before meanComplexity for any given molecule settles into its "consensus" value. Before that the meanComplexity for any given molecule is too dependent on the particular sample of votes, and the meanComplexities are less consistent among molecules.

Figure 7 (bottom) shows the Pearson correlation of meanComplexity for each subset against the final set. Clearly, the CV-R^2 is maximum when the meanComplexity for a subset most closely approximates the meanComplexity for the full set of votes (that is when the Pearson correlation >0.98).

IS THERE A CASE FOR BUILDING DEPARTMENTAL MODELS OF MEANCOMPLEXITY?

One possible situation is that voters in individual departments would have their own set of rules (e.g., process chemists vs analytical chemists vs computational chemists) and that pooling data over all chemists would produce a less self-consistent global model as compared to more self-consistent departmental models. That concern can be addressed by looking at the CV-R^2 of departmental models. A departmental model uses the meanComplexity calculated from votes of chemists in individual departments. Making departmental models, however, was not productive. Individual departments fall on the curve of Figure 7 (top). This implies that individual departments are not significantly more or less self-consistent than a random subset of votes of the same size. Most departmental models have very low CV-R^2 because individual departments have $\ll$30 votes per molecule. One can eliminate the number of molecules as an important factor in the CV-R^2 because all departmental models with more than ~8 votes per molecule include the full set of 2681 molecules, and in most of these the CV-R^2 is clearly poor.

WHAT ARE THE IMPORTANT DETERMINANTS OF MEANCOMPLEXITY?

The descriptor types MOE_2D, SP3CARBONS, and DESCRIPTORCOMPLEXITY together contain 207 individual descriptors. One can query a QSAR model about individual descriptor importances. In the case of RF, this is done by randomly permuting the values of each individual descriptor to the wrong molecules and monitoring how much the out-of-bag predictions degrade. The 10 most important descriptors are shown in Table 2. We see meanComplexity is affected by surprisingly few factors. The number of chiral centers (SP3CARBONS_CHIRAL_COUNT and MOE_2D_CHIRAL, which presumably measure the same thing) is clearly the most important. This is not at all surprising. Many previous models of complexity all include the number of chiral centers. MOE_2D_WEINERPOL is a topological index that measures the sum of interatomic distances in a molecule,[34] which includes a notion of "branching" or "compactness". More branched molecules are subjectively more complex. Finally, DESCRIPTORCOMPLEXITY_UNIQUETT is the number of unique topological torsions in the molecule. That is, the fewer unique local chemical environments in the molecule, the less complex it appears. This is consistent with an observation

dx.doi.org/10.1021/ci5001778 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

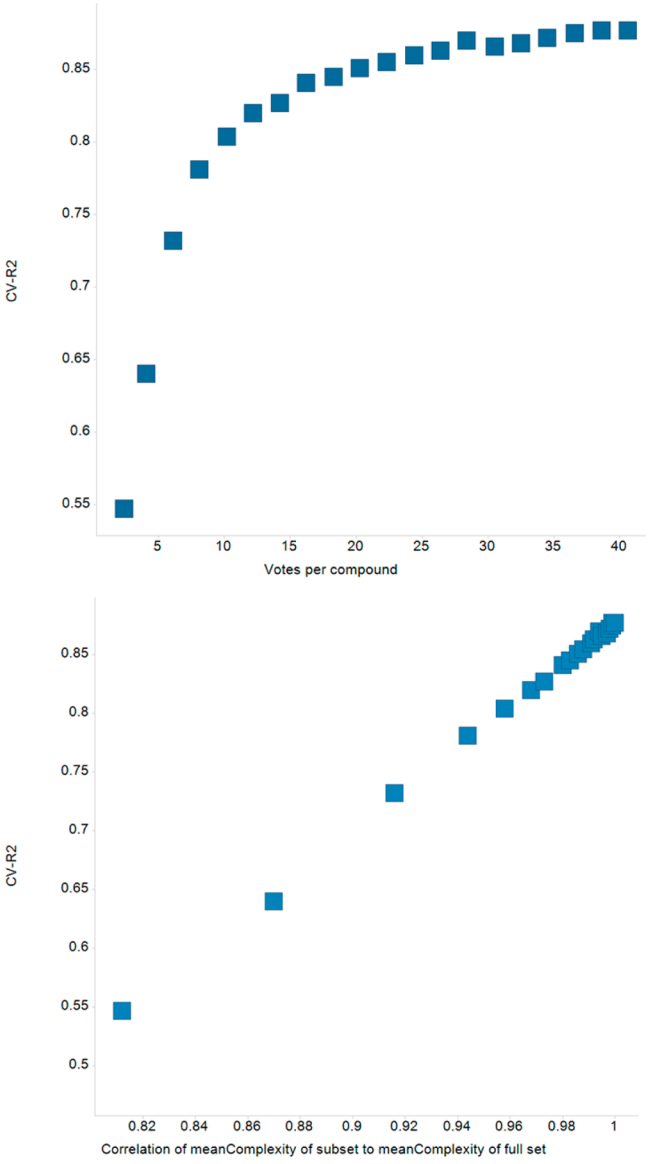


Figure 7. Cross-validated R^2 for randomly selected subsets of votes as a function of votes per compound in the subset (top) and as a function of Pearson correlation of meanComplexity in subset vs meanComplexity in full set (bottom). Each point represents a randomly selected subset of votes.

Table 2. Ten Most Important Descriptors for the RF Model Using MOE_2D, SP3CARBONS, and DESCRIPTORCOMPLEXITY

parameter	value
SP3CARBONS_CHIRAL_COUNT	248
MOE_2D_CHIRAL	247
MOE_2D_WIENERPOL	182
DESCRIPTORCOMPLEXITY_UNIQUETT	170
SP3CARBONS_CHIRAL_ALLCARBON_RATIO	80
MOE_2D_GCUT_SMR_3	78
MOE_2D_ZAGREB	78
SP3CARBONS_CHIRAL_ALLATOM_RATIO	58
MOE_2D_GCUT_PEOE_3	43
DESCRIPTOR_COMPLEXITY_UNIQUE_AP	41

about complexity by von Korff and Sander[14] using a different type of substructure descriptor. It is interesting that there is no important descriptor associated with the overall size of a molecule (number of non-hydrogen atoms, molecular weight, etc.).

The meaning of the three most important descriptors can be appreciated by comparing pairs of compounds that have near-median values of the other three descriptors but differ in the descriptor of interest. Table 3 shows some examples. The pair 662308/1051128 differs in the number of chiral centers. The pair 290554/314374 differs in MOE_2D_WEINERPOL. The first molecule is "unbranched" and "elongated", while the second is "branched" and "compact". The pair 215904/7714 differs in DESCRIPTORCOMPLEXITY_UNIQUETT. The first molecule contains a great deal of internal similarity, i.e. fewer unique TT descriptors. Although it is not an important descriptor for complexity, NONHYDROGEN_ATOMS is included to demonstrate that the important descriptors can vary even if the effective size of the molecule is nearly constant.

Figure 8 shows scatterplots of meanComplexity vs SP3CARBONS_CHIRAL_COUNT, MOE_2D_WEINERPOL, DESCRIPTORCOMPLEXITY_UNIQUETT, and NONHYDROGEN_ATOMS. Although MOE_2D_WEINERPOL and DESCRIPTORCOMPLEXITY_UNIQUETT are correlated with NONHYDROGEN_ATOMS, the first two better correlate with meanComplexity (R^2 = 0.548, 0.579, 0.426, respectively).

A nonlinear model like from RF will clearly fit the data closely, especially since meanComplexity is not linear with some important properties (for example SP3CARBONS_CHIRAL_COUNT in Figure 8). However, it is possible to make a simple, more interpretable model using a few descriptors. For example

$$\begin{aligned}\text{meanComplexity} = {} & -0.989 \\ & + 0.0175\ \text{DESCRIPTORCOMPLEXITY\_UNIQUETT} \\ & + 4.60\ \text{SP3CARBONS\_CHIRAL\_ALLTATOM\_RATIO} \\ & + 0.510\ \text{MOE\_2D\_VDISTMA} \\ & - 0.514\ \text{MOE\_2D\_VDISTEQ}\end{aligned}$$

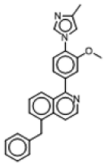
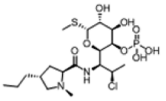
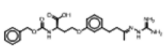
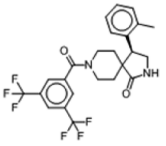
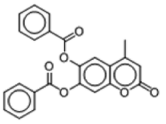
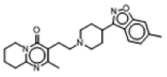
is a linear model with a CV-R^2 of 0.80. We see the same types of properties as in the RF model, although which specific descriptors are most important may vary. DESCRIPTORCOMPLEXITY_UNIQUETT is among the most important descriptors in the RF model. SP3CARBONS_CHIRAL_ALLTATOM_RATIO is the ratio of the number of chiral centers to the total number of non-hydrogen atoms, an alternative way of looking at chirality. MOE_2D_VDISTEQ and MOE_2D_VDISTMA are alternative measures of "compactness".

CAN WE MODEL STDEVCOMPLEXITY?

One common exercise for crowdsourcing is to identify molecules on which voters tend to disagree and then try to discern which features are responsible for the disagreement. The amount of disagreement in our case is quantified by stdevComplexity. The distribution of stdevComplexity is apparent from Figure 2 (bottom). Any modeling of stdevComplexity should have a correction for the downturn near meanComplexity of 1 and 5. Otherwise, low stdevComplexity would correlate with the structural properties of molecules with especially low or high complexity. One can generate a spline or a polynomial equation that would pass through the points in Figure 2 (bottom), including the downturns. The corrected stdevComplexity would be the original stdevComplexity minus the spline value. Positive corrected stdevComplexity would mean voters disagreed about that molecule more than expected; negative corrected stdevComplexity would mean voters agreed about the molecule more than expected.

dx.doi.org/10.1021/ci5001778 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 3. Pairs of Molecules with Disparate Values for Important Descriptors

MOLECULE	NONHYDROGEN_ATOMS	CHIRAL_CENTERS	WEINERPOL	UNIQUETT	STRUCTURE
662308	31	0	51	38	
1051128	31	9	50	40	
290554	33	1	41	34	
314374	34	1	62	40	
215904	30	0	48	20	
7714	30	0	51	54	

Unfortunately, we could not find a self-consistent model of corrected or uncorrected stdevComplexity using any combination of descriptors (CV-R^2 < 0.2). That is, there is no structural feature that would predict whether Merck chemists would agree or disagree about a molecule. We can also show that the number of votes for an individual molecule is not a significant predictor of stdevComplexity.

HOW WELL DOES MEANCOMPLEXITY CORRELATE WITH SYNTHETIC ACCESSIBILITY INDICES IN THE LITERATURE?

While there are many papers about molecular complexity, only a few of them include the structures on which their models are based, so it is difficult to make a direct comparison. In Figure 9 we show two extreme cases. Ertl and Schuffenhauer[9] determined a model for synthetic accessibility (SAscore) of drug-like molecules based on a "fragment score" minus a "complexity penalty". (Higher SAscore actually means lower "accessibility".) The fragment score was based on what fraction of the molecule contained frequently occurring fragments in PubChem. Interestingly, in a small crowdsourcing exercise, chemists scored 40 molecules for their synthetic accessibility and their consensus correlated with the calculated SAscore. Those 40 molecules were included in our set (Source = "Ertl" in Supporting Information). Figure 9 (top) shows the correlation of SAscore with meanComplexity. The agreement (R^2 = 0.89) is very high, slightly better than the internal consistency of our own modeled meanComplexity (CV-R^2 = 0.88). This is remarkable given that the derivation of SAscore is very different from the derivation of meanComplexity and that SAscore is meant to represent synthetic accessibility rather than complexity. It would be more interesting to compare meanComplexity with the complexity penalty from Ertl and Schuffenhauer for those 40 molecules, but that number is not provided in that publication.

At the other extreme, Figure 9 (bottom) shows meanComplexity compared to RSynth, an index of retrosynthetic complexity implemented in the MOE software[26] for all our molecules. Rsynth of "1" means easy to synthesize and "0" means difficult to synthesize. There is no correlation of Rsynth with meanComplexity (R^2 = 0.02). Clearly, different ways of estimating synthetic accessibility may not at all correlate. For example, a molecule that is complex as seen by one method (e.g., with many chiral centers) may appear very synthetically accessible in a retrosynthetic view if most of the chiral centers are contained in a single preexisting reagent.

HOW WELL DOES MEANCOMPLEXITY CORRELATE WITH PMI?

Process mass intensity (PMI) has become an important metric in the field of pharmaceutical process development and can help in ranking the synthetic desirability and environmental impact of different API processes.[10,15] PMI is the mass of all materials consumed divided by the mass of product. The lower the PMI, the more efficient the synthesis. PMI was the topic of a recent modeling paper by Kjell et al.[15] where the authors sought to build

dx.doi.org/10.1021/ci5001778 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

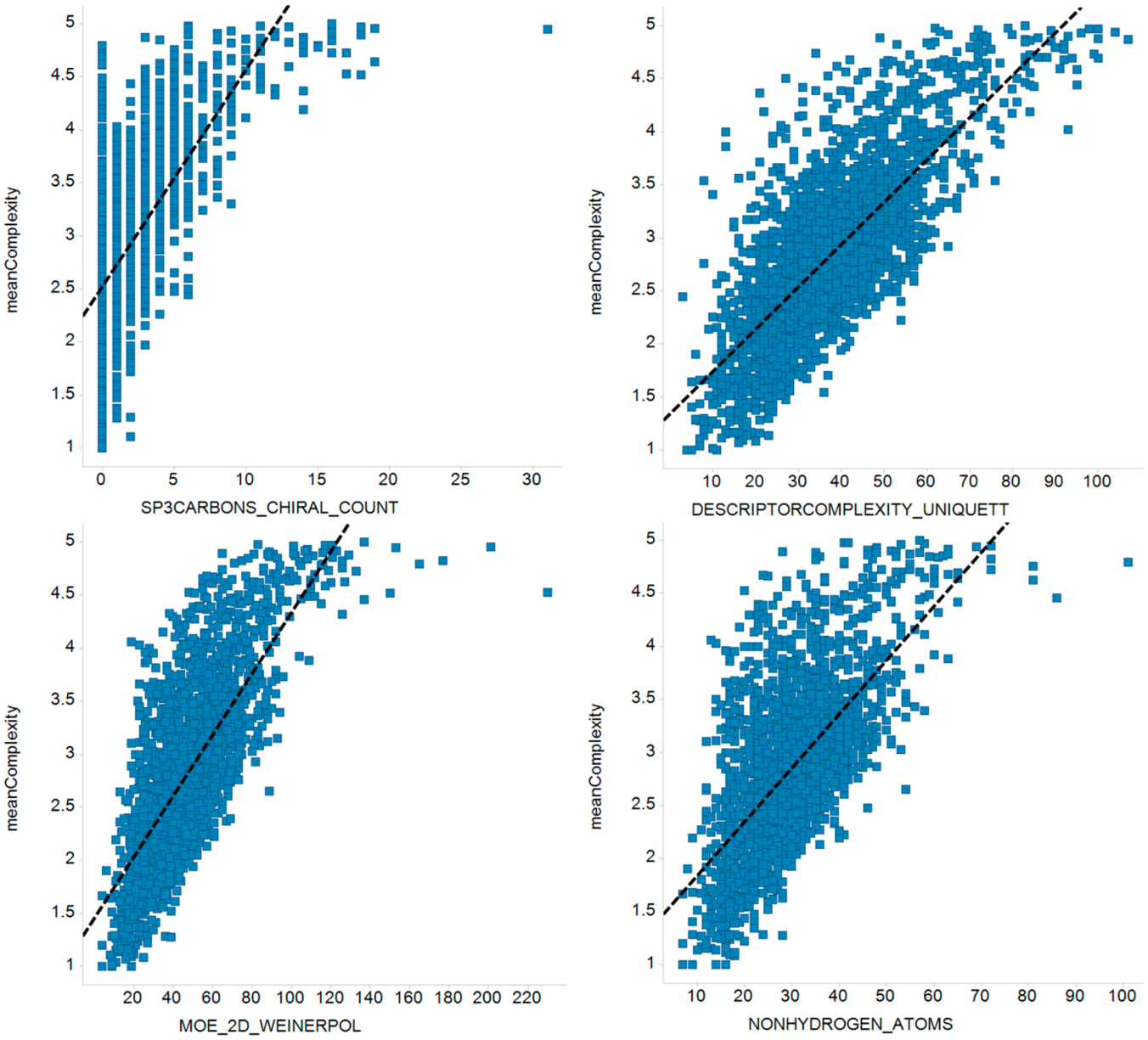


Figure 8. The scatterplots of meanComplexity vs individual descriptors. The dashed line represents the best least-squares line. The R^2 of the best fit relationships are 0.48, 0.55, 0.58, 0.43.

a predictive model of PMI. The training set was a proprietary set of 14 molecules which were then used to predict the experimental PMI for 14 additional molecules (excluding one outlier) which are included in our set (Source = "Kjell" in the Supporting Information). Figure 10 (top) shows the correlation of the measured PMI from Kjell et al. vs meanComplexity. There is a significant correlation (R^2 = 0.58). Interestingly, Kjell et al.'s own model of PMI could not predict the experimental PMI of these molecules (R^2 = 0.05).

Extending the possible utility of a direct correlation between molecular complexity and PMI, we looked internally to a test set of 24 proprietary molecules from the Merck molecule collection with PMI values ranging from 75 to 1120. From Figure 10 (bottom) it is evident that a trend in the predicted PMI with increasing meanComplexity is present. The R^2 of the fit increases from 0.30 to 0.57 if the two molecules with the largest PMI are not included in the analysis, making it close to the correlation for the Kjell et al. PMI. The R^2 of the correlation of PMI against the number of non-hydrogen atoms is 0.29, excluding the same two outliers. We are currently exploring the connection between molecular complexity and PMI in more detail.

DISCUSSION

This paper combines the ideas of molecular complexity and crowdsourcing. There are many suggestions in the literature for how to define complexity. Most investigators select a small number of characteristics that seem intuitively correlated with complexity and combine them using weights. Later, it might be demonstrated that the complexity rule correlates with the subjective assessment of complexity by chemists or to some experimental evaluation of synthetic difficulty. Ultimately, this approach seems too arbitrary since different investigators pick out different characteristics. Just to take one example, Whitlock[3] suggests 4 x number of rings + 2 x number of saturations + 1 x number of heteroatoms + 2 x number of chiral centers. On the other hand, Barone and Chanon[4] would not just consider the

dx.doi.org/10.1021/ci5001778 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

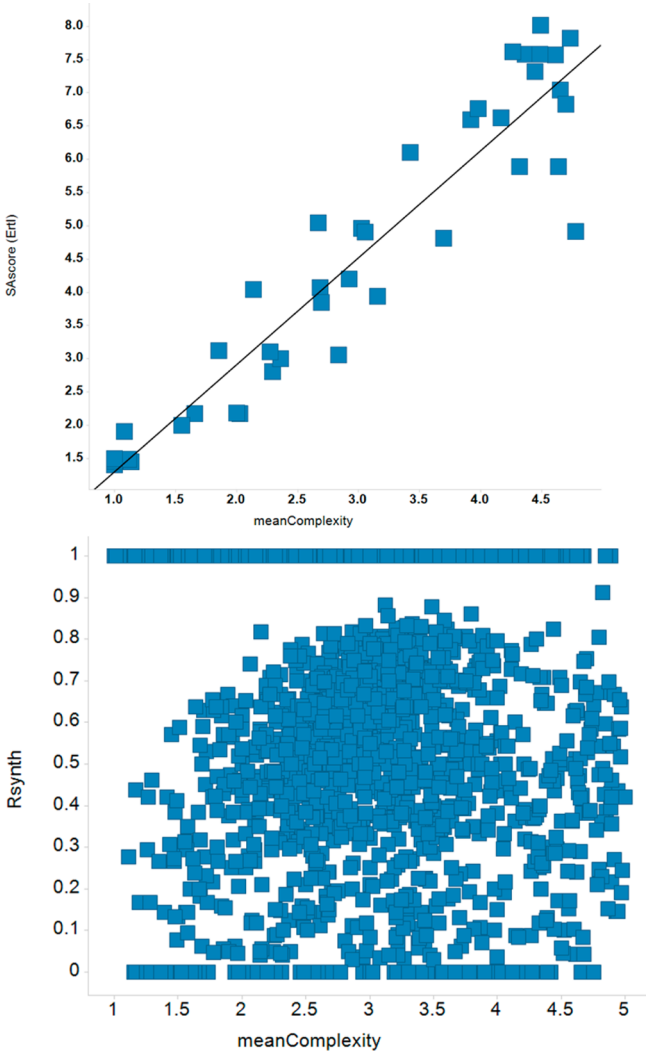


Figure 9. (top) Comparison of SAscore from Ertl and Schuffenhauer to meanComplexity R^2= 0.89 and to Rsynth from MOE R^2 = 0.05.

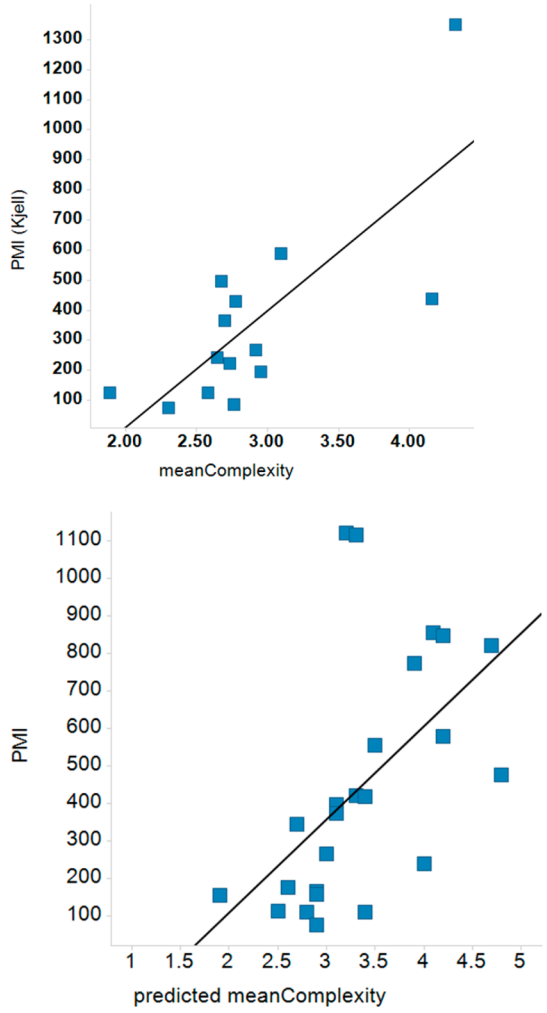


Figure 10. (top) Comparison of measured PMI from Kjell et al. to meanComplexity (R^2 = 0.59). (bottom) Comparison of predicted PMI from Merck to predicted meanComplexity (R^2 = 0.57) omitting the two highest PMI values.

number of rings but the size of the rings. Here we do the reverse: start with the subjective assessment of complexity, derive a model from that, and ask the model what the important characteristics are. In this case the assessment of complexity is derived from in-house crowdsourcing exercise.

In medicinal chemistry crowdsourcing has been applied mostly to whether a molecule is suitable to be acquired or further developed as a lead.[16−18] Ours appears to be the only crowdsourcing exercise specific to complexity. However, crowdsourced estimates of synthetic accessibility are precedented,[9,12] albeit using fewer chemists and fewer molecules than we have here.

The critical assumption behind the "wisdom of crowds" idea is that any given individual may be biased, but that the biases will cancel when many individuals are polled. In this case it appears to be true. In our study chemists generally did not agree with each other about the complexities of individual molecules, but once a consensus meanComplexity was generated by averaging over many voters, it was easy to build a very self-consistent QSAR model of complexity. One lesson here, which is not often addressed elsewhere, is that a minimum average number of votes per compound may be necessary to achieve maximum self-consistency. Another lesson is that, as long as a representative sample of votes is obtained per compound, it is not necessary for every chemist to score every compound.

Now that we have generated a way of defining complexity for any given molecule and have shown that our complexity is at least moderately predictive of synthetic accessibility and/or efficiency, we can imagine the following applications:

1. Monitor the complexity of compounds registered at Merck over time to see whether there have been any changes in the difficulty of molecules being synthesized.
2. Sort in-house therapeutic programs by complexity.
3. Prioritize the purchase of external libraries by complexity.
4. Follow the mean complexity of compounds in a lead optimization program to see if complexity is increasing with time to unacceptable levels.
5. Use complexity as one parameter in a multiparameter optimization.

Figures 11−13 show three examples of these applications. In Figure 11 compounds synthesized at Merck or purchased from commercial sources between 1972 and 2013 are binned by year. The *y*-axis is the average predicted meanComplexity for that year. The average complexity of synthesized compounds (blue) shows a slow rise until the late 1980s, after which the complexity shows small excursions. Generally the complexity of purchased

dx.doi.org/10.1021/ci5001778 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

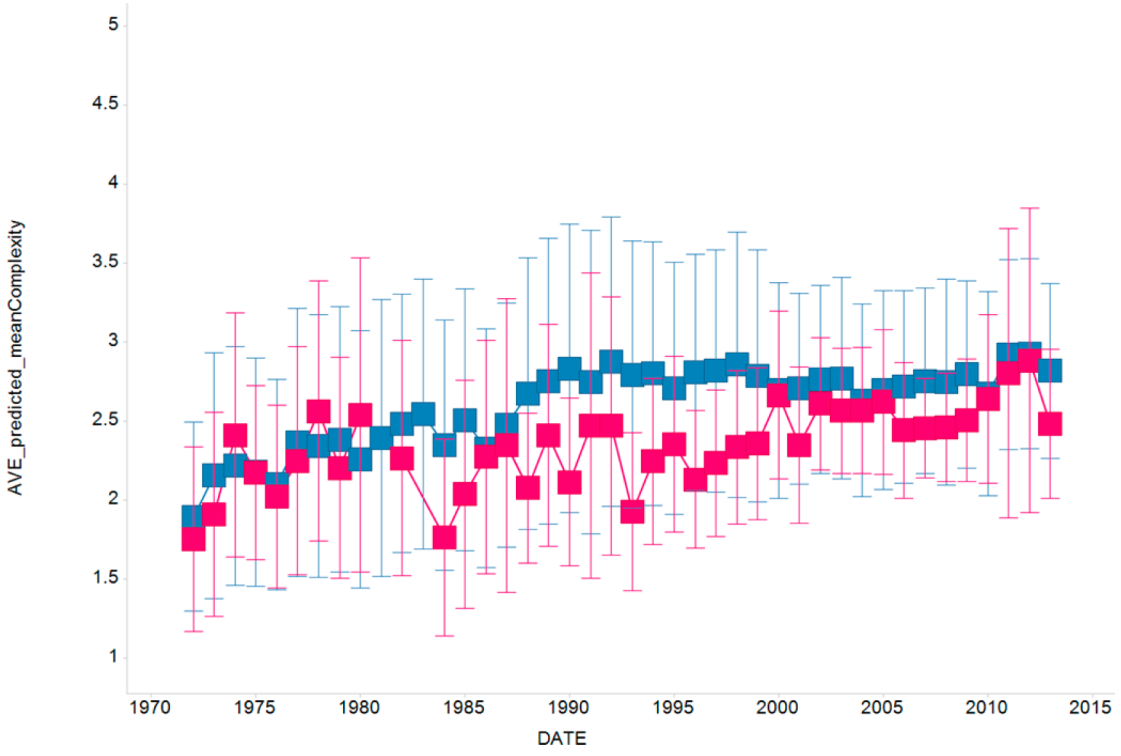


Figure 11. Average ±1 stdev in predicted meanComplexity for compounds synthesized at Merck (blue) or purchased by Merck from commercial sources (red) from 1972 to 2013 binned in yearly intervals.

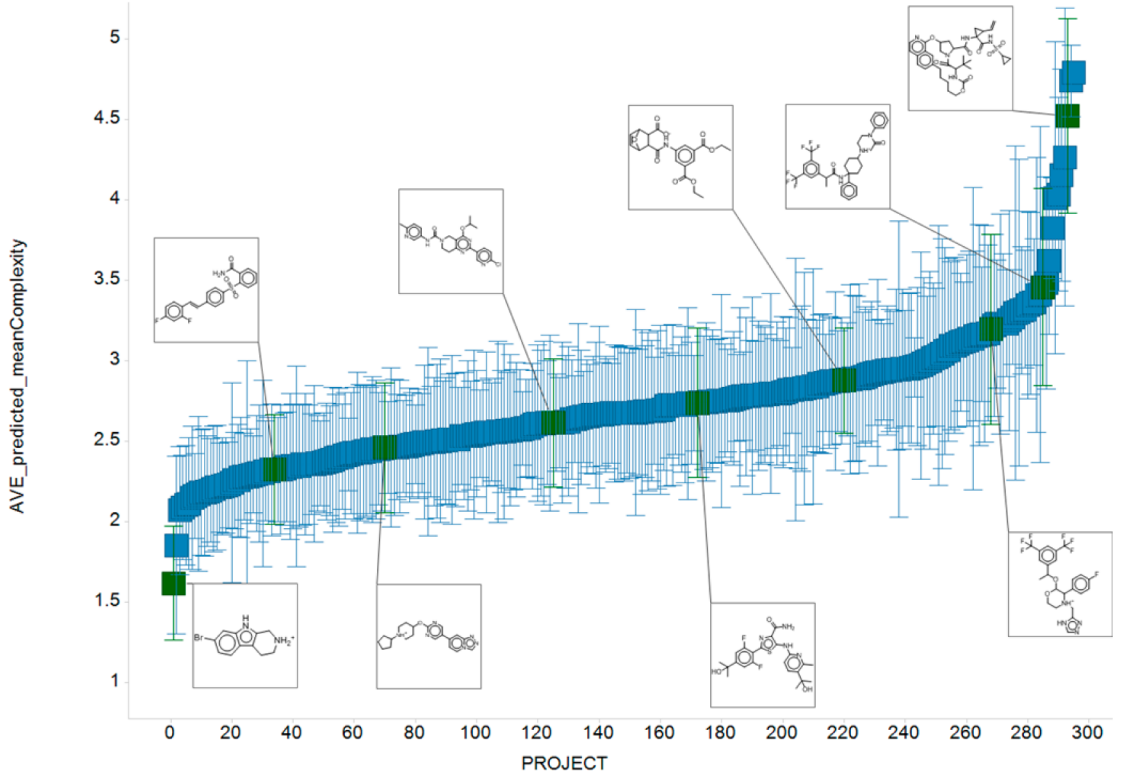


Figure 12. Average ±1 stdev in predicted meanComplexity for ∼300 in-house programs. Published or patented structures with predicted meanComplexity closest to the average are shown for selected programs.

compounds (red) is less than the synthesized compounds, and there is much more variation between years. This reflects whether the goal in any given year was to purchase more drug-like compounds or reagents.

In Figure 12 we are looking at ∼300 in-house synthetic programs sorted by increasing average predicted meanComplexity. For selected programs we are showing the published or patented Merck compounds from that program with a predicted

dx.doi.org/10.1021/ci5001778 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

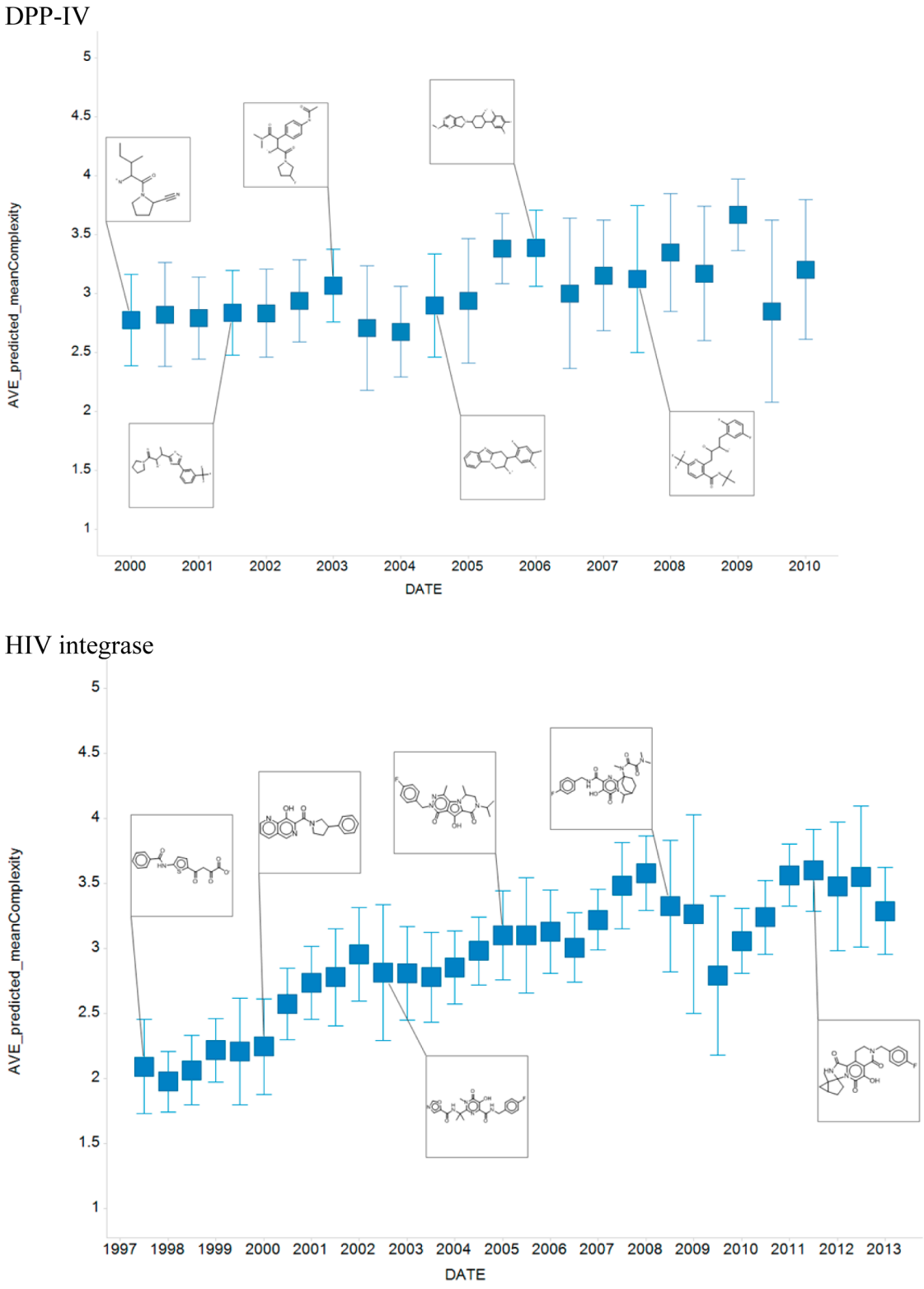


Figure 13. Average ±1 stdev predicted meanComplexity for two in-house programs as a function of time at six-month intervals. Published or patented structures with predicted meanComplexity closest to the average are shown for selected time periods.

meanComplexity closest to the average for the program. Clearly some programs have more complex molecules than others, and that is seen in the structures.

In Figure 13 we monitor the complexity of compounds synthesized for the DPP-IV and HIV integrase programs at six month intervals. DPP-IV is an example where the complexity of compounds is more or less constant with time. HIV integrase is an example where the complexity increases with time.

■ ASSOCIATED CONTENT

Supporting Information

We provide for the nonproprietary molecules in this study: meanComplexity, stdevComplexity, number of votes, SMILES

dx.doi.org/10.1021/ci5001778 | J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

strings, and descriptors. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: sheridan@merck.com.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We are grateful to all members of the Merck community for participating in the scoring exercise. We would like to thank the Merck Scientific Advisory Committee and Chemistry Council for guidance and input. We acknowledge Chris Culberson, Brad Feuston, and Joe Shpungin for scripts used as part of Merck's MIX modeling environment. Chris Culberson helped us extract registration dates and project assignments for in-house compounds. We would like to thank Frank Brown, Chris Culberson, Ian Davies, Mike Hack (Johnson & Johnson), Chris Hill, Andy Liaw, Daniel McMasters, Brad Sherborne, and Chris Waller for helpful comments.

REFERENCES

(1) Bertz, S. H. First general index of molecular complexity. *J. Am. Chem. Soc.* **1981**, *103*, 3599−3601.

(2) Hendrichson, J. B.; Huang, P.; Toczko, A. G. Molecular complexity: a simplified formula adapted to individual atoms. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 63−67.

(3) Whitlock, H. W. On the structure of total synthesis of complex natural products. *J. Org. Chem.* **1998**, *63*, 7982−7989.

(4) Barone, R.; Chanon, M. A new a simple approach to chemical complexity. Application to the synthesis of natural products. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 269−272.

(5) Bertz, S. H. Complexity of synthetic reactions. The use of complexity indices to evaluate reactions, transforms and disconnections. *New J. Chem.* **2003**, *27*, 860−869.

(6) Rucker, C.; Rucker, G.; Bertz, S. H. Organic synthesis—art or science. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 378−386.

(7) Allu, T. K.; Oprea, T. I. Rapid evaluation of synthetic and molecular complexity for in silico chemistry. *J. Chem. Inf. Model.* **2005**, *45*, 1237−1243.

(8) Schuffenhauer, A.; Brown, N.; Selzer, P.; Ertl, P.; Jacoby, E. Relationship between molecular complexity, biological activity, and structural diversity. *J. Chem. Inf. Model.* **2006**, *46*, 525−535.

(9) Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminf.* **2009**, *1*:8.

(10) Jimenez-Gonzalez, C.; Ponder, C. S.; Broxterman, Q. B.; Manley, J. B. Using the right green yardstick: Why process mass intensity is used in the pharmaceutical industry to drive more sustainable processes. *Org. Process Res. Dev.* **2011**, *15*, 912−917.

(11) Leach, A. R.; Hann, M. M. Molecular complexity and fragment-based drug discovery: ten years on. *Curr. Opin. Chem. Biol.* **2011**, *15*, 489−496.

(12) Bonnet, P. Is chemical synthetic accessibility computationally predictable for drug and lead-like molecules? A comparative assessment between medicinal and computational chemists. *Eur. J. Med. Chem.* **2012**, *54*, 679−689.

(13) Soh, S.; Wei, Y.; Kowalczyk, B.; Gothard, C. M.; Baytekin, B.; Gothard, N.; Grzybowski, B. A. Estimating chemical reactivity and cross-influence from collective chemical knowledge. *Chem. Sci.* **2012**, *3*, 1497−1502.

(14) Von Korff, M.; Sander, T. About complexity and self-similarity of chemical structures in drug-discovery. *Chaos and Complex Systems*; Sttavrinides, S. G. et al., Eds.; Springer-Verlag: Berlin, 2013; pp 301−306.

(15) Kjell, D. P.; Watson, I. A.; Wolfe, C. N.; Spitler, J. T. Complexity-based metric for process mass intensity in the pharmaceutical industry. *Org. Process. Res. Dev.* **2013**, *17*, 169−174.

(16) Hack, M. D.; Rassokhin, D. N.; Buyck, C.; Seierstad, M.; Skalkin, A.; ten Holte, P.; Jones, T. K.; Mirzadegan, T.; Agrafiotis, D. K. Library Enhancement through the Wisdom of Crowds. *J. Chem. Inf. Model.* **2011**, *51*, 3275−3286.

(17) Kutchukian, P. S.; Vasilyeva, N. Y.; Xu, J.; Lindvall, M. K.; Dillon, M. P.; Glick, M.; Coley, J. D.; Brooijmans, N. Inside the mind of a medicinal chemist: the role of human bias in compound prioritization during drug discovery. *PLoS One* **2012**, *11*, e48476.

(18) Peng, Z.; Gillespie, P.; Weisel, M.; So, S.-S.; So, W. V.; Kondru, R.; Narayanan, A.; Hermann, J. C. A crowd-based process and tool for HTS triage. *Mol. Inf.* **2013**, *32*, 337−345.

(19) MDDR reference http://accelrys.com/products/databases/bioactivity/mddr.html (accessed May 2, 2014).

(20) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100−D1107.

(21) 200 Top Selling Drugs for 2010. www.drugs.com (accessed May 2, 2014).

(22) Irwin, J. J.; Shoichet, B. K. ZINC − A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177−182.

(23) Pipeline Pilot reference http://accelrys.com/products/pipeline-pilot/ (accessed May 2, 2014).

(24) Carhart, R. E.; Smith, D. H.; Ventkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and application. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.

(25) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological torsions: a new molecular descriptor for SAR applications. comparison with other descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82−85.

(26) Molecular Operating Environment (MOE), Version 2008, release 10, Chemical Computing Group, Montreal, Canada, 2009. http://www.chemcomp.com/ (accessed May 2, 2014).

(27) Breiman, L. Random forests. *Machine Learning* **2001**, *45*, 5−32.

(28) 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.

(29) Cortes, C.; Vapnik, V. N. Support-Vector Networks. *Machine Learning* **1995**, *20*, 273−297.

(30) LIBLINEAR -- A Library for Large Linear Classification http://www.csie.ntu.edu.tw/~cjlin/liblinear/ (accessed May 2, 2014).

(31) Svetnik, V.; Wang, T.; Tong, C.; Liaw, A.; Sheridan, R. P.; Song, Q. Boosting: an ensemble learning tool for compound classification and QSAR modeling. *J. Chem. Inf. Model.* **2005**, *45*, 786−799.

(32) Bruce, C. L.; Melville, J. L.; Picket, S. D.; Hirst, J. D. Contemporary QSAR classifiers compared. *J. Chem. Inf. Model.* **2007**, *47*, 219−227.

(33) 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.

(34) Balaban, A. T. Five new toplogical indices for the branching of tree-like graphs. *Theor. Chim. Acta* **1979**, *53*, 355−375.