Supporting Information

A Data-Driven Strategy for Predicting Greenness Scores, Rationally Comparing Synthetic Routes and Benchmarking PMI Outcomes for the Synthesis of Molecules in the Pharmaceutical Industry

Jun Li, Eric M. Simmons and Martin D. Eastgate*

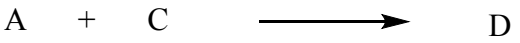
Chemical and Synthetic Development, Bristol-Myers Squibb, 1 Squibb Drive, New Brunswick, NJ, 08903 (USA)

E-mail: martin.eastgate@bms.com

Table of Contents

Derivatization of the relationship between Cumulative and Step PMI	**PageS2-S3**
Actual step PMI in JAK2, Brivanib, CCR2, and Apixaban Processes	**PageS4-S6**
Cumulative PMI linear synthesis demo	**PageS7-S10**
Cumulative PMI convergent synthesis demo	**PageS10-S11**
Relationship between step PMI and molar step yield	**PageS12-S14**
Caveat and future direction	**PageS14-S17**

The cumulative PMI at step *i* in above linear synthesis is determined by the cumulative PMI at step *i-1* as shown in eqn (1). To find the relationship between the cumulative PMI and the step PMI as defined in eqn (2), one can combine both eqn (1) and eqn (2), followed by converting the mass ratio of the substrate vs. the product into the step molar yield and molecular weights to obtain the eqn (3)



A: Substrate (Prepared from previous step or purchased)
C: Non-substrates (incl. reagents, solvents for reaction and workup, catalyst, aqueous solution, filtering aid and etc)
D: Isolated product

Scheme S1: A simple linear synthesis example at step *i*

$$CumulativePMI_i = \frac{Mass_A \times CumulativePMI_{i-1} + Mass_C}{Mass_D} \quad (1)$$

$$stepPMI_i = \frac{Mass_A + Mass_C}{Mass_D} \quad (2)$$

$$CumulativePMI_i = \frac{Mw_A}{Mw_D \times Yield_i}(CumulativePMI_{i-1} - 1) + stepPMI_i \quad (3)$$

A: Substrate A (Prepared from previous step or purchased) (limiting)
B: Substrate B (Prepared from previous step or purchased) (n equiv)
C: Non-substrates (incl. reagents, solvents for reaction and workup, catalyst, aqueous solution, filtering aid and etc)
D: Isolated product

Scheme S2: A simple convergent synthesis example at step *i*

$$CumulativePMI_i = \frac{Mass_A \times CumulativePMI_{i-1} + Mass_B \times CumulativePMI_j + Mass_C}{Mass_D} \quad (4)$$

$$stepPMI_i = \frac{Mass_A + Mass_B + Mass_C}{Mass_D} \quad (5)$$

$$CumulativePMI_i = \frac{Mw_A}{Mw_D \times Yield_i}(CumulativePMI_{i-1} - 1) + \frac{n \times Mw_B}{Mw_D \times Yield_i}(CumulativePMI_j - 1) + stepPMI_i \quad (6)$$

Actual step PMI in JAK2, Brivanib, CCR2, and Apixaban Processes

The process mass intensity for each step in JAK2 synthesis was determined after the scale-up from an initial clinical campaign (**Table S1**), and the cumulative PMI for the API was about 3600.

Steps	Step PMI	Reaction Type
1⇨2	18	Bromination
2⇨3	30	Ester formation
3⇨4	138	Telescope (Alkylation-Borylation)
4⇨5	100	Suzuki
5⇨6	26	Amidine Formation
6⇨7	73	Cyclization
7⇨8	222	Ulmann Coupling
8⇨9	47	Saponification
9⇨10	129	Amidation
S1-1⇨S1-2	31	Condensation
S1-2⇨S1-3	214	Cyclization
S1-3⇨S1-4	16	Bromination
S2-1⇨S2-2	152	Telescope (Bromination-Cyclization)
S2-2⇨S2-3	161	Sandmeyer
S3-1⇨S3-2	35	Alkylation
S3-2⇨S3-3	70	Deallylation
S3-3⇨S3-4	52	Reductive amination
S3-4⇨S3-5	91	Deallylation
Cumulative	3600	

Table S1. Actual step PMI observed in JAK2 synthetic sequence on-scale

The process mass intensity for each step in Brivanib synthesis was determined from a late stage LTSS campaign (**Table S2**), and the cumulative PMI for the API was about 1488.

Steps	Step PMI	Reaction Type
11⇨12	22	enamine formation
12⇨13	48	Cyclocondensation
13⇨15	98	Telescope (N-amination-Cyclocondensation)
15⇨16	63	Grignard addition
16⇨18	113	Telescope (Continuous oxidation-ester formation)
18⇨20	79	Telescope (Chlorination-SNAr)
20⇨22	44	Telescope (Saponification-epoxide ring opening)
22⇨24	91	Telescope(Ester formation-Hydrogenolysis)
S1-5⇨S1-6	76	Telescope(Ester formation-deBoc)
S2-4⇨S2-6	34	Telescope (SNAr-Decarboxylation)
S2-6⇨S2-7	22	SNAr
S2-7⇨S2-8	35	Deprotection
S2-8⇨S2-9	64	Indole formation
Cumulative	1488	

Table S2 Actual step PMI observed in Brivanib synthetic sequence on-scale

The process mass intensity for each step in CCR2 synthesis was determined from an early stage clinical campaign (**Table S3**), and the cumulative PMI for the API was about 1717.

The process mass intensity for each step in Apixaban synthesis was determined from a late stage validation campaign (**Table S4**), and the cumulative PMI for the API was about 197. To illustrate the percentage of the contribution from each of the step in the cumulative PMI, we showed that the penultimate step (43→44) had the largest impact with approximate 28% of the cumulative PMI (**Figure S1**).

Steps	Step PMI	Reaction Type
25⇨26	26	Claisen condensation
26⇨27	8	enamine formation
27⇨28	46	enamine reduction
28⇨29	13	Hydrogenolysis
29⇨30	23	Amidation
30⇨31	26	Methylation
31⇨32	82	Cycloalkylation
32⇨33	62	ketal deprotection
33⇨34	28	reductive amination
34⇨35	92	Saponification
35⇨36	65	Curtis Rearrangement
36⇨37	53	Telescope(Deboc-Acylation)
37⇨38	29	Hydrogenolysis
38⇨39	210	Telescope(Chlorination-SNAr)
S1-7⇨S1-8	10	Iodination
S1-8⇨S1-9	70	Pd-cat Cynation
S1-9⇨S1-10	53	Cyclocondensation
Cumulative	1717	

Table S3 Actual step PMI observed in CCR2 synthetic sequence on-scale

Steps	Step PMI	Reaction Type
40⇨41	30.6	Telescope (Amidation-Cycloalkylation-Chlorination)
41⇨42	21	Elimination
42⇨43	15.2	Cycloaddition
43⇨44	47.6	Telescope (Nitroreduction-Amidation-Cycloalkylation)
44⇨45	49.6	Amidation
S1-12⇨S1-13	25.3	Condensation
Cumulative	197	

Table S4. Actual step PMI observed in optimized Apixaban synthetic sequence on-scale

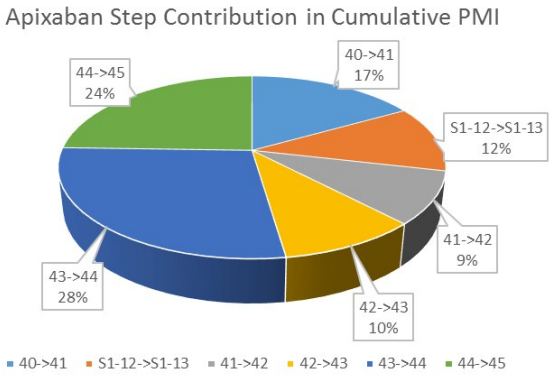
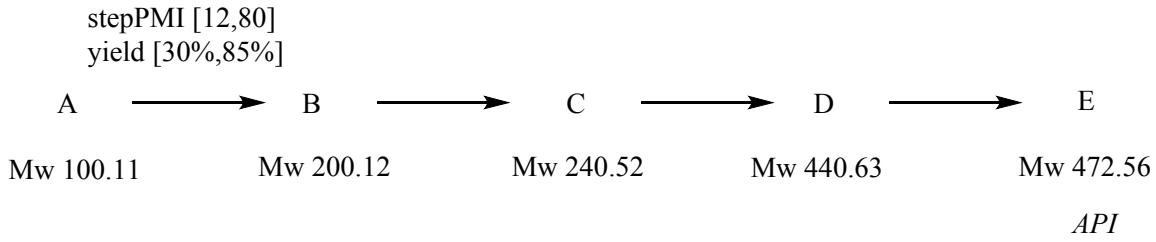


Figure S1. Step contribution in cumulative PMI in Apixaban synthesis

Cumulative PMI linear synthesis demo

The Monte Carlo simulation of a cumulative PMI for a linear synthesis is demonstrated using the R statistical programming language[1] below. In the following demo for a hypothetical 4-step linear synthesis as shown in **Scheme S3**:



Scheme S3

We assume that step PMI and step molar yield can be sampled from a negatively correlated bivariate normal distribution.[2] The assumption allows us to provide input ranges between a pair of optimistic and pessimistic values for both step PMI and yield for each of the steps within the synthetic sequence. These ranges can be estimated based on the existing data in the database, other published scale-up procedures for a similar chemical transformation, or experiences acquired by the chemists during the studies. For example in the A to B chemical scale-up process, a pair of negatively correlated step PMI and step yield can be sampled with anticipated step PMI between 12 to 80, and the anticipated molar step yield between 30 to 85 M%. The ranges can be based on distributions of the step PMIs and yields observed for the similar transformations on scale from the green chemistry database. Typically, we start to select the ranges between minimum and maximum of the observed distributions unless some of the examples are deemed as the outliers, therefore excluded. This can be done by narrowing down the range to between 1st and 3rd quantiles. Conceptually, step PMI is governed by type of

chemical transformations, the ensuing work-up conditions, along with the solubility of molecules. When we segment the chemistry transformation types for step PMI, other factors including the solubility were confounded. In theory, the more compounds are included in the green chemistry database, the wider the solubility span of the compounds could be, and the more evenly distributed of the solubility span could be among different transformations. This will ultimately help average out the solubility impact during the segmentation. We typically used the step PMI ranges without corrections from solubility unless we have additional working experiences on the solubility of the intermediates which are on the extreme ends of the spectrum and not captured in the proposed transformation type. Considering the possibility of employing extractive work-up processes, we may adjust the upper bound of the step PMI accordingly (for example 20~50% increase is anticipated).

Based on the cumulative PMI formula, a recursive function can be used to efficiently obtain the sequence of the cumulative PMI along the linear synthetic branch. To transform the pair of optimistic and pessimistic range (high-low) into the standard deviation (sd) input for the bivariate normal distribution function, we can derive the sd through dividing the range (high-low) by 5.15 to capture the 99% sampling between the pair. Similarly, one can use either 3.29 or 4.0 instead of 5.15 to obtain 90% or 95% PI respectively. The R codes for the correlated bivariate normal distribution function and cumulative PMI recursive function for the linear branch are shown in the **Figure S2**.

Once, we obtain the predictive distribution of the cumulative PMI scores of the synthesis from the Monte Carlo simulation, we can i) compare different proposed synthetic routes in terms of their predicted mean and 90% prediction intervals; ii)

compare the actual cumulative PMI score if available from unoptimized processes with the predicted ranges. As shown in the **Figure S3**, the actual cumulative PMI score of 280 ranked it against the predicted cumulative PMI histogram, showing it is better than 77% of the predicted scores of similar transformations on scale (red section), while it is still worse than 23% of the predicted scores (black section).

```r
#Correlated bivariate normal distribution function
bicor <- function (n, pmi_mean, pmi_sd, yield_mean, yield_sd, correlation) {
  #n:           number of sample
  #pmi_mean:    estimated PMI average
  #pmi_sd:      estimated PMI standard deviation
  #yield_mean:  estimated molar yield average
  #yield_sd:    estimated molar yield standard deviation
  #correlation: estimated correlation between PMI and yield
  x <- rnorm(10000)
  y <- rnorm(10000)
  z <- correlation * scale(x)[,1] + sqrt(1 - correlation^2) * scale(resid(lm(y ~ x)))[,1]
  x_new <-  pmi_mean + pmi_sd * scale(x)[,1]
  y_new <-  yield_mean + yield_sd * z
  d     <-  data.frame(x=x_new,y=y_new)
  idx   <-  sample(1:nrow(d), n)
  return (d[idx,])
}

#Cumulative PMI Recursive function for the linear branch
br1.cum.pmi<-function(i){
  if (i==1) return (br1.step.pmi[i])
  else return (br1.SM[i]/br1.Pdt[i]/br1.Yield[i]*(br1.cum.pmi(i-1)-1)+br1.step.pmi[i])
}
```

Figure S2. Snippets of R code used for cumulative PMI recursive function and bivariate normal distribution.

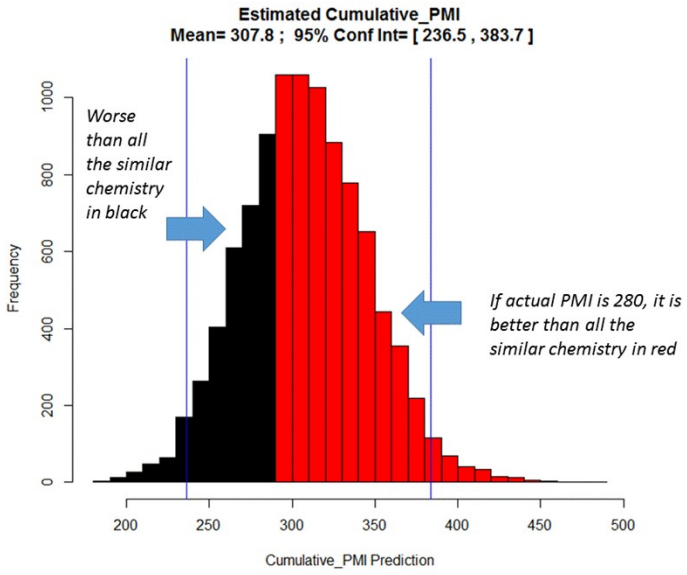
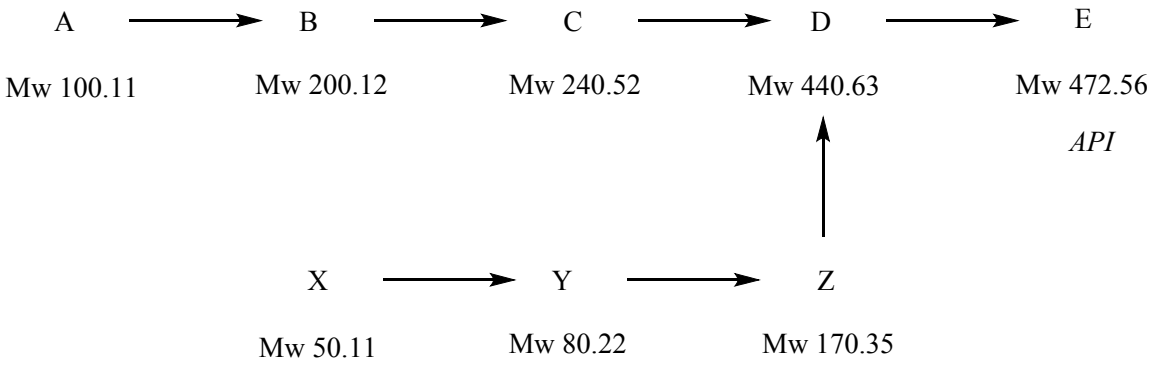


Figure S3 Interpretation of observed PMI in the predictive distribution of the cumulative PMI from the linear synthesis.

Cumulative PMI convergent synthesis demo



Scheme S4

For the convergent synthesis, a hypothetic synthetic **Scheme S4** was demonstrated here. The coding blocks shown here are not optimized (**Figure S4**). Basically, it breaks down the above scheme into three linear branch syntheses (A to C, X to Z, and D to E) which used all the cumulative PMI recursive functions.

```
#Branch 1 from A to C
br1.cum.pmi<-function(i){
  if (i==1) return (br1.step.pmi[i])
  else return (br1.SM[i]/br1.Pdt[i]/br1.Yield[i]*(br1.cum.pmi(i-1)-1)+br1.step.pmi[i])
}
#Branch 2 from X to Z
br2.cum.pmi<-function(i){
  if (i==1) return (br2.step.pmi[i])
  else return (br2.SM[i]/br2.Pdt[i]/br2.Yield[i]*(br2.cum.pmi(i-1)-1)+br2.step.pmi[i])
}

#Branch 3 from D to E
br3.cum.pmi<-function(i,bif_cum_pmi1){
  if (i==1) return (br3.SM[i]/br3.Pdt[i]/br3.Yield[i]*(bif_cum_pmi1-1)+br3.step.pmi[i])
  else return (br3.SM[i]/br3.Pdt[i]/br3.Yield[i]*(br3.cum.pmi(i-1,bif_cum_pmi1)-1)+br3.step.pmi[i])
}
```

Figure S4. Snippets of R code used for cumulative PMI recursive functions in convergent synthesis

The coupling reaction block (C+Z to D) was demonstrated below. Typically, for prediction purpose, we assume equimolar condition for the coupling reaction between two branches unless one of the branches must be using large excess based on existing knowledge and experiences.

```
#Bifurcate function
#Note: Here for prediction purpose, the assumpition was always using equimolar condition
#for sake of simplicity
#In reality, it may need to factor in "n equivalents" to match the actual condition!
#If using n, make sure CumPMI1 and Mw1 always the limiting agent.
bif.cum.pmi<-function(CumPMI1,CumPMI2,n,stepPMI,Y,Mw1,Mw2,MwP){
  return (Mw1/MwP/Y*(CumPMI1-1)+n*Mw2/MwP/Y*(CumPMI2-1)+stepPMI)
}
```

Figure S5. Snippets of R code used for coupling reaction in convergent synthesis

All the R codes for linear and convergent synthesis demos discussed in the supplementary materials are available upon request from the authors.

Relationship between step PMI and molar step yield

A general reaction can be represented as shown in **Scheme S5,** where the starting material reacts with n equivalent of reactants/reagents. We denote C as the total other input masses combined, including all solvents, aqueous media, catalysts, ligands, inorganics, filter aids, etc. Based on the step PMI eqn (7) and molar step yield, we can derive the relationship through eqn (9). $\frac{Mass_C}{Mass_{SM}}$, the ratio of $Mass_C$(all other input masses combined) to $Mass_{SM}$ (starting material mass) can be represented by S, arriving at eqn (10). Basically S is defined as the unit mass, indicating the kilograms of all nonconsequential input masses (not participating in apparent bond-forming/breaking of the desired product) per kilogram of starting material. Further derivatization from eqn (10) through eqn (12), we obtained the term $MW_{SM} + n\sum Mw_R - Mw_P$, which essentially describes all the combined molecular weights of the side products generated between the starting material and the reactants/reagents in the absence of excess reactants/reagents if *n* is equimolar to the starting material, for example the urea side product generated in a typical EDAc coupling.

Evidently, we can see that S can be a predominant factor in the case of i) dilute reactions such as ring closing metathesis, macrolactamization, or very low substrate solubility ii) chromatography, iii) multiple acid/base washes iv) scrubbing tanks to remove harmful volatiles ranging from HCN, hydrazoic acid, methyl iodide, methanethiol etc. On the other hand, the contribution from the side products are related to the type of chemical transformations and the ratio of the molecular weight of side-products to the molecular weight of isolated product. In case of large excess of

reactants/reagents, unconsumed reagents can be treated as contributing to the term S. Generally speaking, if product molecular weight is relatively small, the use of chemistry such as Mitsunobu or a Simmons-Smith cyclopropanation, which generates relatively high molecular weight of the side products, can become a major factor in increasing step PMI.

SM: Starting material (limiting)
R: Reactants/reagents (n equiv)
C: Other masses combined (incl. solvents for reaction and workup, catalyst, aqueous solution, filtering aid and etc)
P: Isolated product

Scheme S5. General reaction scheme

$$stepPMI = \frac{Mass_{SM} + Mass_R + Mass_C}{Mass_P} \tag{7}$$

$$stepPMI = \frac{Mass_{SM} + \frac{Mass_{SM}}{Mw_{SM}} \times n\sum Mw_R + Mass_C}{\frac{Mass_{SM}}{Mw_{SM}} \times Mw_P \times stepYield} \tag{8}$$

$$stepPMI = \frac{1 + \frac{1}{Mw_{SM}} \times n\sum Mw_R + \frac{Mass_C}{Mass_{SM}}}{\frac{1}{Mw_{SM}} \times Mw_P \times stepYield} \tag{9}$$

$$stepPMI = \frac{1}{stepYield} \times \frac{Mw_{SM} + n\sum Mw_R + S \times Mw_{SM}}{Mw_P} \tag{10}$$

$$stepPMI = \frac{1}{stepYield} \times \frac{Mw_{SM} + n\sum Mw_R - Mw_P + Mw_P + S \times Mw_{SM}}{Mw_P}$$

(11)

$$stepPMI = \frac{1}{stepYield} \times \left(\frac{Mw_{SM}}{Mw_P} \times S + 1 + \frac{Mw_{SM} + n\sum Mw_R - Mw_P}{Mw_P} \right) \quad (12)$$

In principle, one could use eqn (12) with an estimated yield range as an input for a particular step, along with a range of S as well as molecular weights of side-products based on the chemistry type chosen and standard processes in the scale-up, to predict the step PMI, and subsequently propagate through a proposed synthetic sequence to obtain a predicted cumulative PMI. Although this is conceptually feasible, the estimation on the ranges of S and side-products can be inconvenient due to these type of statistics not being readily available. For the proof of concept, the negatively correlated bivariate distribution in both step PMI and yield was simulated using a bivariate normal distribution.

Caveat and future direction

It should be noted that negatively correlated bivariate normal model is a crude simplification of the complex relationship between step PMI and molar step yield. Here the differences in distribution between a bivariate model (53% negatively correlated based on the overall data in the green chemistry database) and above derived model (eqn 6) were compared under the assumptions in eqn 6: i) ratio of $\frac{Mw_{SM}}{Mw_P}$, follows a uniform distribution between 0.8 and 1.2; ii) ratio of $\frac{MW_{SM} + n\sum Mw_R - Mw_P}{Mw_P}$ follows a uniform

distribution between 0.1 and 0.4; iii) S follows a normal distribution centered around either 20 or 28 with standard deviation of 4.

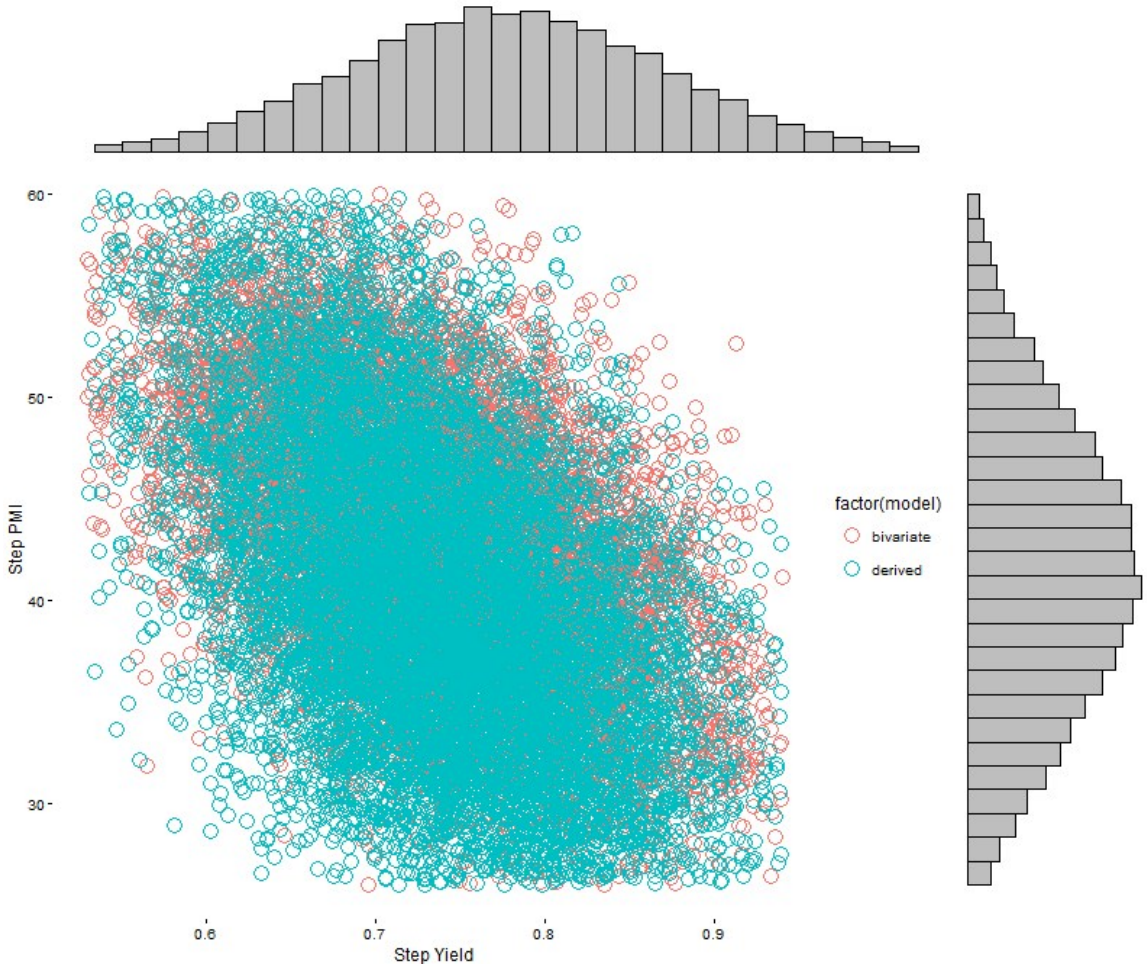


Figure S6. Comparison of correlated bivariate normal model to the derived model (eqn 6) in which S follows a normal distribution centered at 28

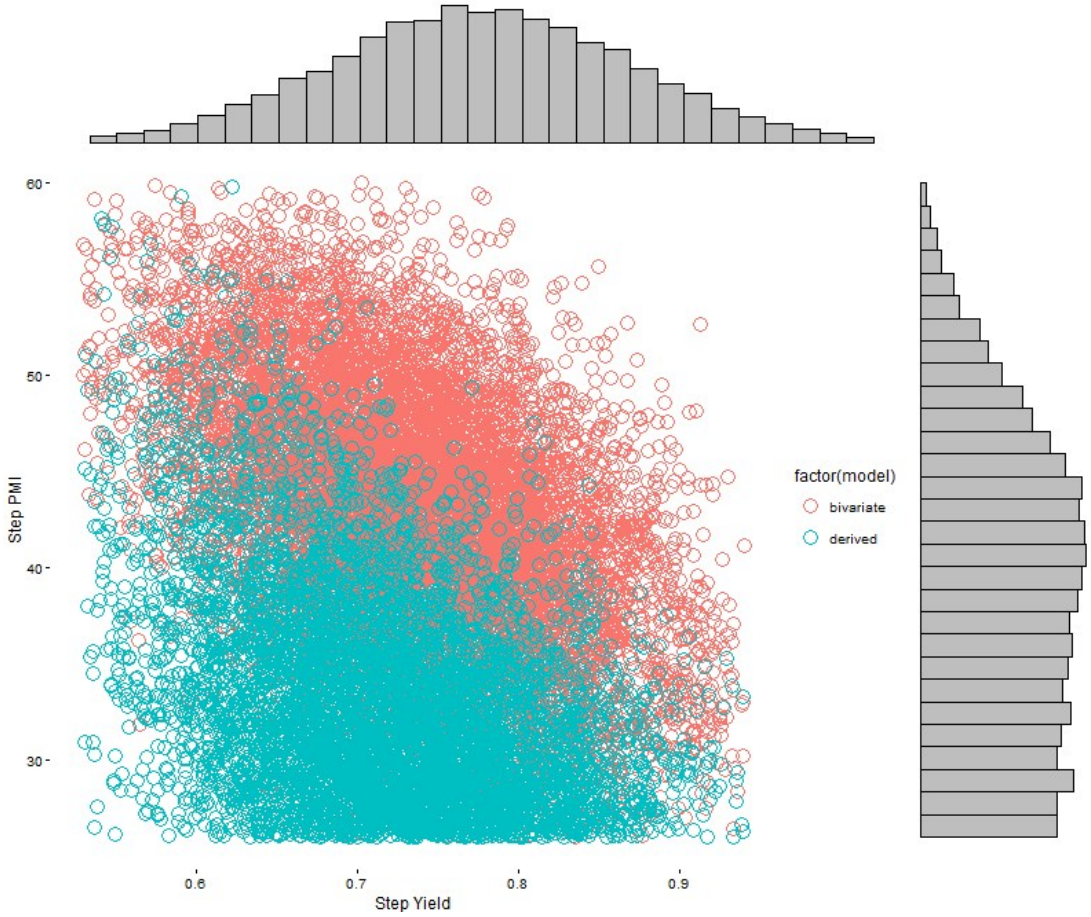


Figure S7. Comparison of correlated bivariate normal model to the derived model (eqn 6) with S centered at 20.

S, the kilograms of all nonconsequential input masses (not participating in apparent bond-forming/breaking of the desired product) per kilogram of starting material, includes all the solvents used in reaction, extractive workup, crystallization or column purifications, all the chemicals and aqueous solutions used for quenching, washing or scrubbing purposes of the processes, all the filtering aids or column medium for color, metal, impurities removal and etc. If S is at high end, which approximately covers the range of 8 to 48 kg per kilogram of starting material, the corresponding step PMI distribution matched pretty well with the bivariate normal model (**Figure S6)**. However,

if S is at relatively low end of the spectrum, which covers the range of 1.6 to 38 kg per kilogram of starting material, the bivariate normal model can underrepresent the low step PMI region (**Figure S7)**.

Based on the studies above, for comparison of the relative synthetic efficiency between different proposed synthetic routes in case of decision-making, this predictive modeling approach with simplified step PMI and yield bivariate normal function provided us with reasonable accuracy for example in the JAK2 routes predictions as shown in the POC. In case of benchmarking, we anticipate the inclusion of more real world pharma data from industrial processes as well as the development of data collection strategy to support the use of derived model such as eqn 12 should further improve the validity and accuracy of the model in the future.

---

[1] R Core Team 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, http://www.R-project.org/

[2] For simulation with correlated non-normal multivariate distributions, Copulas can be used to build these type of dependence between the random variables.