

Mixture Models in the Exploration of Structure-Activity Relationships in Drug Design

Susan Paddock, Duke University
Mike West, Duke University
S. Stanley Young, Glaxo Wellcome Inc.
Merlise Clyde, Duke University

ABSTRACT

We report on a study of mixture modeling problems arising in the assessment of chemical structure-activity relationships in drug design and discovery. Pharmaceutical research laboratories developing test compounds for screening synthesize many related candidate compounds by linking together collections of basic molecular building blocks, known as monomers. These compounds are tested for biological activity, feeding in to screening for further analysis and drug design. The tests also provide data relating compound activity to chemical properties and aspects of the structure of associated monomers, and our focus here is studying such relationships as an aid to future monomer selection. The level of chemical activity of compounds is based on the geometry of chemical binding of test compounds to target binding sites on receptor compounds, but the screening tests are unable to identify binding configurations. Hence potentially critical covariate information is missing as a natural latent variable. Resulting statistical models are then mixed with respect to such missing information, so complicating data analysis and inference. This paper reports on a study of a two-monomer, two-binding site framework and associated data. We build structured mixture models that mix linear regression models, predicting chemical effectiveness, with respect to site-binding selection mechanisms. We discuss aspects of modeling and analysis, including problems and pitfalls, and describe results of analyses of a simulated and real data set. In modeling real data, we are led into critical model extensions that introduce hierarchical random effects components to adequately capture heterogeneities in both the site binding mechanisms and in the resulting levels of effectiveness of compounds once bound. Comments on current and potential future directions conclude the report.

1 Introduction and Background

Medicinal chemists involved in drug design and discovery synthesize large numbers of pharmacological compounds for initial testing to screen for chemical effectiveness in a laboratory setting. This high-throughput screening process assesses many related candidate compounds, each being created by combining basic building blocks drawn from separate sets of complex molecules, or *monomers*. This activity is called combinatorial synthesis. The screening process aims to identify biologically active compounds using aspects of molecular structure associated with biological activity and effectiveness. Our focus here is on statistical modeling and inferential issues arising in exploring such *structure - activity relationships*.

We consider the simplest case of test compounds created by linking together two monomers drawn from two separate test sets. This produces a *library* of two-monomer compounds; the notation $A + B \rightarrow AB$ identifies monomer A from the first set and monomer B from the second set. Various properties of A and B , such as **molecular weight**, partly characterize the resulting compound AB . Chemical activity is measured through experiments that estimate the extent to which the compound inhibits certain chemical reactions; the resulting outcomes are measures of *potency* of the test compound. The synthesized compounds are designed to bind to a structure of interest; for example, synthesized compounds could bind to healthy body tissue to block natural but harmful compounds from binding to the tissue. The site at which the compounds bind to a cell in the tissue is called a receptor. In addition to the effects of constituent monomers, the outcome potency depends on the *binding configuration* of the compound; that is, the compound may bind to a receptor in such a way that individual binding sites are matched by monomers. Hence an effective bind of a two-monomer compound involves two binding sites at the receptor, with one monomer binding to each site. Given a binding configuration, the chemical/physical monomer characteristics are determinants of potency. In practice, it is impossible to observe or measure binding configuration, though it is understood that the various monomer characteristics may play a role in determining configuration.

Though the perspective of the medicinal chemist is the synthesis of monomers, the relationship of a compound to its binding sites depends little upon how the compound was or will be made. Rather, the binding of a compound is governed by the properties of the receptor sites and the compound itself. With a receptor having two binding sites, S_1 and S_2 , the binding can be AB to S_1S_2 or BA to S_1S_2 ; see Figure 1. A data set with n compounds can then be arranged in 2^n configurations so that, as n increases, the number of possible configurations is enormous. For more complicated problems with more than two sites/two monomers, the situation becomes even more complex, of course. Here we develop and fully explore the two sites/two monomers setup, and discuss models we have de-

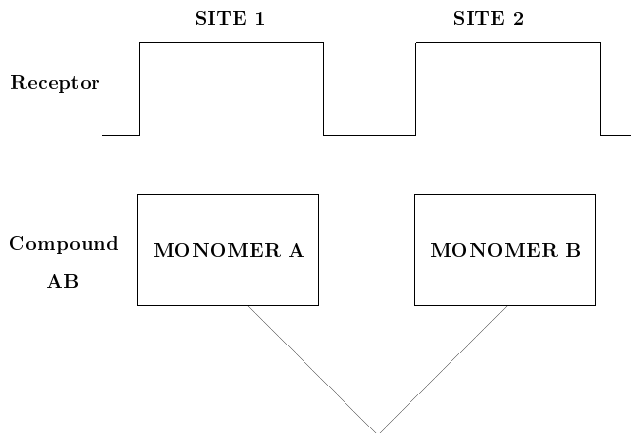


FIGURE 1. Two-monomer, two-site configuration framework. Which orientation is correct: $AB \rightarrow S_1S_2$ or $AB \rightarrow S_2S_1$?

veloped that combine regression models for both binding configurations and potency levels given binding configurations. This introduces two novelties from the chemists' viewpoints: the use of descriptors to model both orientation and binding via separate model components, and the accounting for differences in potency due to binding orientations. This leads to structured mixture models in which linear regressions for potencies are mixed with respect to binary regressions predicting binding configurations. Our work on the development of MCMC methods to implement these models, problems encountered en route, benchmark assessments on simulated data and some preliminary results with real data are described and illustrated here. Our basic mixture model performs well on simulated test data, but is found to be less adequate when applied to real data sets from experiments at Glaxo Wellcome Inc. This leads us into practically critical model extensions that introduce hierarchical random effects components to adequately capture two kinds of heterogeneity in the compound population: significant over-dispersion relative to standard probit predictors in the site binding/orientation mechanism, coupled with a lesser degree of over-dispersion relative to normal linear models in the potency outcome mechanism. We discuss this, explaining the model extensions and developments involved in analysis, with illustration in analysis of a Glaxo Wellcome Inc. data set.

2 Basic Mixture Model and Analysis

2.1 Model Structure

In a context with a two-site receptor S_1S_2 , consider a sample of n two-monomer compounds A_jB_k , ($j = 1, \dots, n_A, k = 1, \dots, n_B$), where n_A and n_B are the number of monomers in sets A and B, respectively, and the resulting $n_A \times n_B = n$ compounds are indexed by i ($i = 1, \dots, n$). Assume that compound binding configurations and potency outcomes are independent across compounds. Covariates are measured on each monomer and our potency model assumes that, given a particular binding configuration, outcomes follow a normal linear regression on the covariates. To account for the two possible binding configurations, we introduce orientation indicators $z_i = 1(0)$ if binding to S_1S_2 is AB (BA). Conditional on binding configurations z_i , the model for outcome potencies y_i is $N(y_i|x_i(z_i)'\beta, \sigma^2)$ where $x_i(z_i)$ is the column vector of monomer covariates, arranged to reflect the configuration determined by z_i . For example, suppose that we measure only the molecular weights of each monomer, w_i^A and w_i^B on monomers A and B , respectively. Then we will take $x_i(1) = (1, w_i^A, w_i^B)'$, corresponding to $AB \rightarrow S_1S_2$, and $x_i(0) = (1, w_i^B, w_i^A)'$, corresponding to $BA \rightarrow S_1S_2$. Generally, $x_i(1)$ is specified with the monomer-specific covariates in a given order, and then $x_i(0)$ contains the same covariates but with the positions of A, B -specific covariates switched. This ensures an unambiguous interpretation of the regression parameters as *site-specific*. In the above example, $\beta = (\beta_0, \beta_1, \beta_2)'$ where β_0 is an intercept term and, for $j = 1, 2$, β_j is the regression coefficient on molecular weight of the monomer that binds to S_j , irrespective of whether that is monomer A or B . The quantity σ measures unexplained dispersion, including assay variability attributable to biological variability present in the receptor and compounds during experimentation and the calibration stage. Nominal levels of assay variability alone are in the 10-20% range on the standard “percent inhibition” scale for potency levels.

Configurations are assumed related to monomer characteristics via a binary regression model. We have adopted a probit link in our work to date, so that the indicators z_i are governed by conditional binding probabilities $\pi_i = Pr(z_i = 1) = \Phi(h_i'\theta)$ where h_i is a column vector of covariates drawn from the same set as those in $x_i(\cdot)$, and $\Phi(\cdot)$ is the standard normal cdf. Hence the model for observed potency outcomes is a mixture over the unobserved configuration, given by

$$y_i \sim \pi_i N(y_i|x_i(1)'\beta, \sigma^2) + (1 - \pi_i) N(y_i|x_i(0)'\beta, \sigma^2)$$

Naturally we deal with this mixture model by explicitly including the latent

indicators z_i . Thus the full model is defined by the joint distribution

$$p(y, z|\beta, \sigma, \theta) = \prod_{i=1}^n N(y_i|x_i(z_i)'\beta, \sigma^2)\Phi(h_i'\theta)^{z_i}(1 - \Phi(h_i'\theta))^{1-z_i}$$

where y and z are the vectors of y_i and z_i respectively. We are interested in inference about the full set of uncertain quantities $(z, \beta, \sigma, \theta)$ based on observing y . Our analyses use standard, conditionally conjugate prior distributions for the regression model parameters, i.e., independent normal priors for β and θ , and inverse gamma priors for σ^2 .

We note that there are symmetries in the model induced by the arbitrary labeling of receptor sites. This is an example of identification features common to most mixture models (Titterton et al. 1985, West 1997). Note that

$$p(y, z|\beta, \sigma, \theta) = p(y, 1 - z|\beta^*, \sigma, -\theta)$$

where β^* is a permutation of the elements of β obtained by switching the labels of receptor sites S_1 and S_2 . While predictions are unaffected by non-identifiability, synthesis of future compounds based upon posterior inferences of properties of A and B is problematic in such a non-identified model. Below we note this and describe how to handle the issue *a posteriori* in the context of MCMC analysis.

2.2 Posterior computations

We exploit the standard latent variable augmentation method for probit models (Albert and Chib 1993) to enable direct Gibbs sampling (Gelfand and Smith 1990) for posterior computation. Introduce latent variables ω_i such that $\omega_i|\theta \sim N(h_i'\theta, 1)$, implying that $\pi_i = Pr[\omega_i \geq 0]$. Then the set of full conditional posterior distributions has the following structure:

- The posterior for $\beta|z, y, \sigma$ is a normal distribution resulting from the conditional linear regression of the y_i on $x_i(z_i)$.
- The posterior for $\sigma^2|z, y, \beta$ is inverse gamma, also a direct derivation from the conditional linear regression of the y_i on $x_i(z_i)$.
- The posterior for $\theta|w$ is a normal distribution resulting from linear regression of the ω_i on h_i .
- The ω_i are conditionally independent and $\omega_i|z_i, \theta$ has a truncated normal posterior (following Albert and Chib 1993).
- The z_i are conditionally independent with posterior odds on $z_i = 1$ versus $z_i = 0$ given by

$$\frac{\Phi(h_i'\theta)N(y_i|x_i(1)'\beta, \sigma^2)}{\{(1 - \Phi(h_i'\theta))N(y_i|x_i(0)'\beta, \sigma^2)\}}$$

The calculation and simulation of each of these distributions, in turn, is standard and easy, and so enables direct Gibbs sampling. The question of parameter identification is dealt with in the posterior computations by examining MCMC samples from an identified region of the parameter space, rather than by attempting to restrict the simulation analysis to such a region by imposing constraints in the prior. This follows what is now essentially standard practice in other kinds of mixture models (e.g., West 1997). Finally, in order to explore predictive questions, for both model evaluation (as in Gelman et al. 1996) and for use in predicting potency of new compounds, we simulate predictive distributions based on the posterior parameters and latent variable samples.

3 A Simulated Dataset

We report an analysis of a test dataset simulated to explore and validate the model and our implementation, which is illuminating in its own right. Here we generated $n = 200$ potency values under a model with one covariate on each monomer. The covariate values are sampled from the $U(-5, 5)$ distribution for monomer A , and from the standard normal distribution for monomer B . Binding orientations z_i are generated from the probit regression on the two covariates with regression vector $\theta = (-0.5, -0.25, -0.5)'$. Then, given orientations, potency levels are drawn from the normal linear regressions with parameter vector $\beta = (30, 3, 8)'$ and variance $\sigma^2 = 9$. The values here were chosen to produce observed data configurations resembling real data sets, although the residual variance σ^2 is lower than typical levels of assay variability. Our analysis is based on quite vague proper priors in the conditionally conjugate class, namely $\sigma^{-2} \sim Ga(\sigma^{-2}|.01, 1)$, $\beta \sim N(\beta|0, 10000I)$, and $\theta \sim N(\theta|0, I)$. Our MCMC analyses have involved a range of experiments with Monte Carlo sample sizes and starting values, and MCMC diagnostics. Following this, our summary numerical and graphical inferences here are based on post burn-in samples of size 20,000. The resulting histogram approximations to marginal posteriors for parameters appear in Figure 2, with related posterior inferences in Figure 3. The top row in this latter graph displays posteriors of the (posterior) classification probabilities of orientation for four selected observations. The actual orientation probabilities and indicators are $\{.67, .68, .45, .86\}$ and $\{0, 1, 0, 1\}$, respectively. Note the concordance of posteriors and true values. To provide insight into aspects of model adequacy, posterior predictive checks are examined. Approximate posterior predictive distributions for potency levels of the four example compounds are displayed in Figure 3 (bottom row). The distributions are centered close to the actual observations, which are marked by crosses along the axes, supportive of model adequacy. A further posterior predictive check involves graphical comparison of the actual data

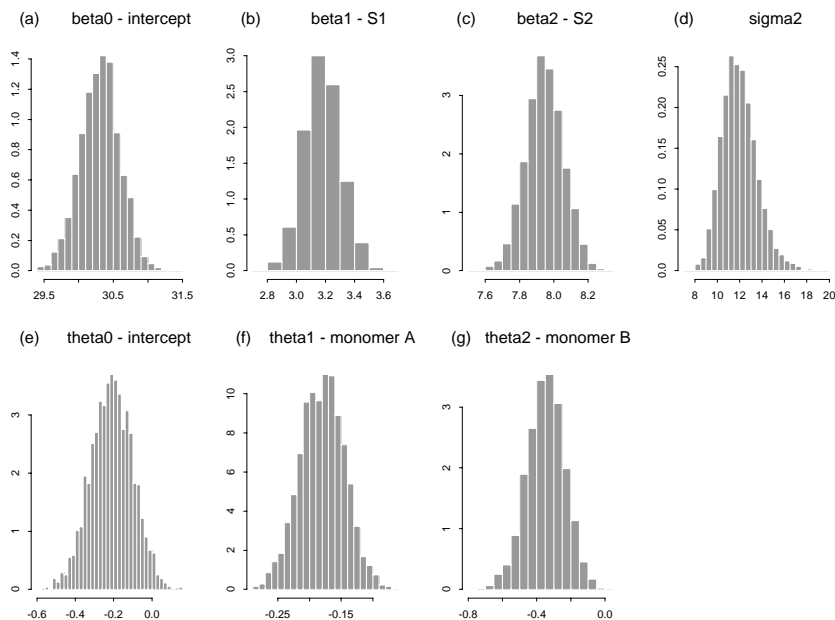


FIGURE 2. (a)-(c) Posteriors for $(\beta_0, \beta_1, \beta_2)$, (d) σ^2 , and (e)-(g) $(\theta_0, \theta_1, \theta_2)$ in analysis of simulated data.

with sampled predictives for the full data set. For a set of draws from the posterior, we sampled 430 observations, one at each design point, producing replicates from the posterior predictive distribution of the data set. Smoothed versions of ten of the resulting histograms are overlaid on the real data histogram in Figure 4. These evidence a comforting degree of conformity to the data histogram.

To the best of our knowledge, we do not encounter the identifiability problem in this analysis. To explore this, we ran another analysis of this simulated data in which the prior for σ^2 was inappropriately biased and concentrated on very large values, namely $\sigma^{-2} \sim Ga(\sigma^{-2}|5, 30000)$. The analysis details were otherwise the same as above. The MCMC trajectories (not shown) from this analysis indicate the lack of identifiability induced by arbitrary labeling of sites; the trajectories for β_1 and β_2 are interchanged two or three times, and the values of the θ_j change sign at the same points. To handle the identification issue we can simply examine a subset of iterations corresponding to one identified region of the parameter space, or map the MCMC samples to a single identified region of the parameter space. The latter strategy is easily implemented in these models by changing the signs of the θ values corresponding to $\theta_1 < 0$, permuting the site-specific β terms at the corresponding iterates, and reflecting the corresponding indicators z_i to $1 - z_i$.

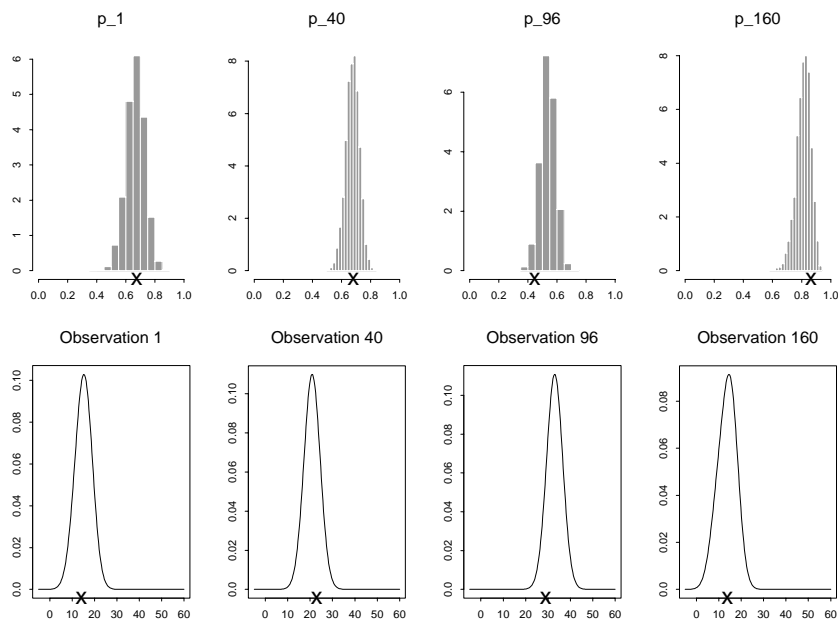


FIGURE 3. *Top row*: Posteriors for classification probabilities of four selected observations in the simulated dataset. *Bottom row*: Posterior predictive densities for the four cases. True values are marked as crosses on the axes in each row.

In summary, this simulated test data analysis indicates that the complicated mixture structure of these models can be adequately unwrapped to identify the effects of differences between monomer characteristics on both binding orientation and potency outcomes.

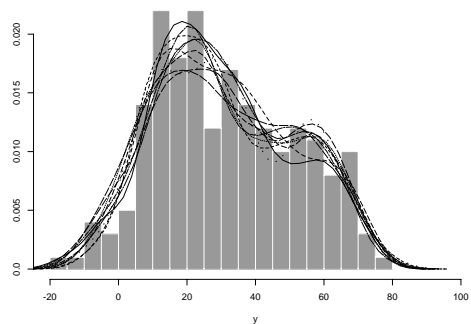


FIGURE 4. Posterior predictive model check for the simulated dataset.

4 Critical Model Elaborations

Initial exploration of the modeling approach on real data sets from laboratories at Glaxo Wellcome Inc. indicated lack of model fit that we address here through model extensions. Deficiencies in the basic model are apparent in that posterior predictive distributions are rather overdispersed, and also obviously biased in some respects, relative to the observed data histogram. Related to this, the corresponding posterior for σ favors values very significantly larger than the ranges of 10-20% for assay standard deviations, which are normally expected by the chemists for assays of this type. We address this via model elaborations that introduce compound-specific random effects to account for and isolate departures from the basic model. The first model extension adds random effects to the orientation/site binding mechanism to relax the strict probit regression on the specific linear predictor chosen. This results in a hierarchical model for over-dispersion in binary regression relative to the probit model. The second extension similarly modifies the strict normality assumption in the linear regressions, inducing heavier tails in the potency outcome distribution. Again this is done by adding hierarchical random effects to allow for additional, compound-specific variability. These two model elaborations very substantially improve model fit to the real data sets, with the resulting interpretations that: (a) there are a number of potency outcomes that are extreme relative to the basic normal regression (conditional on orientation), and (b) there is a significant degree of additional variability in the selection of binding orientations that is simply not captured by the specified linear predictor in the specific probit regressions for binding.

The extended model has the following structure. First, the binding indicators z_i are remodeled via $\pi_i = Pr(z_i = 1) = \Phi(h_i'\theta + \alpha_i)$ where we introduce conditionally independent random effects $\alpha_i \sim N(\alpha_i|0, \tau^2)$. These α_i represent compound-specific effects on binding that are unexplained by the chosen regressors h_i . Inference on these and τ will measure the levels and nature of the extra-probit variability. Second, conditional on binding configurations, the potency outcome levels are remodeled as $y_i \sim N(y_i|x_i(z_i)'\beta, \sigma^2/\lambda_i)$ where we introduce conditionally independent random effects $\lambda_i \sim Ga(\lambda_i|k/2, k/2)$. These λ_i represent compound-specific effects on potency that are unexplained by the chosen regressors $x_i(\cdot)$. This converts the normal error structure of the regression to the more heavy-tailed Student T_k forms induced by integrating out the λ_i (e.g., West 1984).

Application now includes the sets of random effects $\alpha = \{\alpha_i\}$ and $\lambda = \{\lambda_i\}$, together with the hyperparameter τ^2 , in the posterior analysis. We choose to specify the degrees of freedom parameter k but require inference on τ^2 . Details of model modifications are straightforward. Conditional upon $\{\alpha, \lambda\}$, the analysis is changed in only minor ways, adjusting the probit linear predictor and latent variables by the specific α_i , and weighting observations in the linear models by the λ_i . Then, at each stage of the MCMC

analysis, all original model parameters, latent configuration indicators and so forth are sampled from these modified conditional posteriors. This is coupled with resampling of the new random effects and hyperparameters from their corresponding conditionals. These are trivial: the conditional posteriors for the α_i are independent normals, those for the λ_i are independent gammas, and that for τ^2 is an inverse gamma under an inverse gamma prior. Further details can be easily derived by the reader, or can be obtained on request from the authors.

5 A Glaxo Wellcome Dataset

A real data set from Glaxo Wellcome Inc. was selected from an original symmetric three-position library of compounds – that is, a library of three-monomer compounds. We selected 430 compounds with a common monomer and treat this smaller data set as a two-position library. The specific library under study is comprised of compounds similar to those synthesized by Whitten et al. (1996), who examine inhibition for a set of compounds consisting of a central triazine ring with two distinct monomers. Three covariates named `clogP`, `flexibility` and `molecular weight` are given for each monomer; `clogP` is a continuous-valued measure of lipophilicity and `flexibility` and `molecular weight` are integer-valued binned variables. Potency levels y_i measure percent inhibition, which is a measure of how well the synthesized compounds perform with respect to a standard; for example, the standard can be a natural compound for which the medicinal chemists hope to discover a synthetic substitute with similar chemical behavior. At this stage of experimentation, the chemists search for highly potent compounds that will be retested in subsequent experiments to further assess their potential usefulness in drug design.

All three covariates are used in the linear regression model for potency outcomes, so that each x_i is a 7-vector, including the intercept constant. For the orientation model we difference the observed covariate values so that each h_i is a 3-vector representing relative covariates between the two monomers, plus an intercept. Our analysis is based on quite vague proper priors in the conditionally conjugate class. The prior for the illustrated analysis has the following independent margins: $N(\theta|0, I)$, $N(\beta|0, 10000I)$, $Ga(\sigma^{-2}|0.01, 1)$, and $Ga(\tau^{-2}|2, 1)$; we additionally set $k = 10$. Again following experiments with ranges of MC sample sizes and starting values for the MCMC analysis, summary numerical and graphical inferences here are based on a post burn-in sample of 5,000 iterates subsampled from a longer run, and deemed adequate based on repeat runs. We note that there is no appearance of the switching phenomenon in the MCMC trajectories induced by the model identification problem illustrated with our modified simulated data analysis earlier. Though this does not fully assure us that

the MCMC run has remained in a single identified region of the parameter space, it does give some reason to believe that to be the case. Resulting sampled values are therefore assumed to be drawn from a single identified region, and are presented as histograms.

Figure 5(a-c) presents posteriors for the potency outcome regression parameter β . Note the clearly identified differences between the estimated effects of monomer characteristics when binding at different sites. Consider the variable `clogP`, for example, and the posterior in Figure 5(a). This indicates that a unit increase in `clogP` of the monomer at S_1 is associated with an average increase of about 3 units in potency. By contrast, a unit `clogP` increase of the monomer at S_2 has an associated average potency increase of around 17 units. From Figure 5(b) and (c), it appears that the coefficients of `flexibility` and `molecular weight` for the monomer at S_1 are apparently small or negligible, in distinct contrast to those of the monomer at S_2 . All variables are clearly relevant, as is expected by the medicinal chemists, though these differential effects due to binding orientations have never before been identified. The posterior for σ^2 in Figure 5(g) indicates a standard deviation around 15-19, which is large on this outcome scale though quite consistent with the experimental chemistry involved. As noted earlier, assay variability alone is typically benchmarked at levels consistent with σ values around 10-20% on the inhibition scale. This supports the view that the random effects components α_i and λ_i have adequately catered for over-dispersion and mis-specification in the regression components.

Graphs in Figure 5(d-f,h) display the posteriors for the components of θ . It is apparent that the relative values of `clogP` and `molecular weight`, describing the difference between the two monomers, are strongly associated with orientation outcomes, whereas `flexibility` is evidently much less important if not quite irrelevant. As mentioned previously, `clogP` is continuous, whereas the other covariates were computed as binned values; such binning could induce a loss of information resulting in the suppression of a relationship. Nevertheless, these preliminary conclusions are quite consistent with the consensus of chemists, who agree that `clogP` is generally one of the best chemical descriptors available for these data. Our isolation and estimation of the differential effects of `clogP` and `molecular weight` on binding versus potency outcomes is quite novel, and provides more incisive information about the relationships.

The posterior for the probit random effects variance τ^2 , Figure 5(i), supports standard deviation values in the 0.6-0.7 range. This translates to the random effects alone accounting for a substantial range of variability on the orientation probability scale, potentially determining the orientation in some instances. Though the probit regression does isolate a meaningful explanatory predictor, there is evidently a fair degree of residual, compound-specific variation unexplained and that has here been simply estimated, compound by compound, via the α_i random effects. Further work with

chemists should consider examination of compounds with significant random effects to enquire about possible additional explanatory variables and the reliability of assays.

Some global aspects of model adequacy are explored via posterior predictive checks, as illustrated in Figure 5(j,k). For each observed compound, we compute approximate posterior predictive distributions at the observed design points and evaluate the resulting set of 430 predictive quantiles at the actually observed potency outcomes. Assuming the effects of dependencies induced by posterior parameter uncertainties to be ignorable, departures of this set of quantiles from an appearance of approximate uniformity will indicate global model deficiencies. These figures display a uniform quantile plot and a histogram of the values, and the concordance with uniformity is comforting, suggestive of an adequate model fit.

We further explore model fit and interpretation of the random effects parameters in Figure 6. Potency predictions for eight randomly selected compounds appear in the top two rows. Predictions for compounds with extreme posterior means for their α_i and λ_i values appear in rows 3 and 4, respectively. In each case, the posterior means of λ_i and α_i are given, and the observed potency values are marked on the axis. In the randomly selected cases, predictions are made apparently adequately. The compounds in row 3 have apparently extreme α_i values but moderate λ_i values, while the situation is reversed for compounds in row 4. The observed values for the compounds in rows 3 and 4 consistently fall in the tails of their predictive distributions. To explore the implications of accounting for heterogeneity of the compounds with respect to activity, consider the first histogram in row 4. The mean of the λ_i here is 0.292. The observed potency outcome is -29 , which is initially puzzling since the data measure percent inhibition. Despite the assumption of inhibition for the compounds under experimentation, it is the case that excitation occurs occasionally due to biological variability, explaining this outcome that is quite outside the model for inhibition we have constructed. Hence the posterior indicates a very low λ_i value, isolating this case as an apparent outlier. Here, as in the other cases, it is interesting to note that lack of fit of the original model is isolated via extreme values of just one of the two random effects, indicating that compounds poorly represented by the basic model depart in terms of either extreme potency outcomes, or in terms of an unusual binding orientation, but, at least for the selected compounds, not both simultaneously.

6 Discussion

Our analysis, a new approach to assessing chemical structure-activity relationships, promises to be useful for examining libraries of compounds through exploration of monomer space. Our results also provide motivation

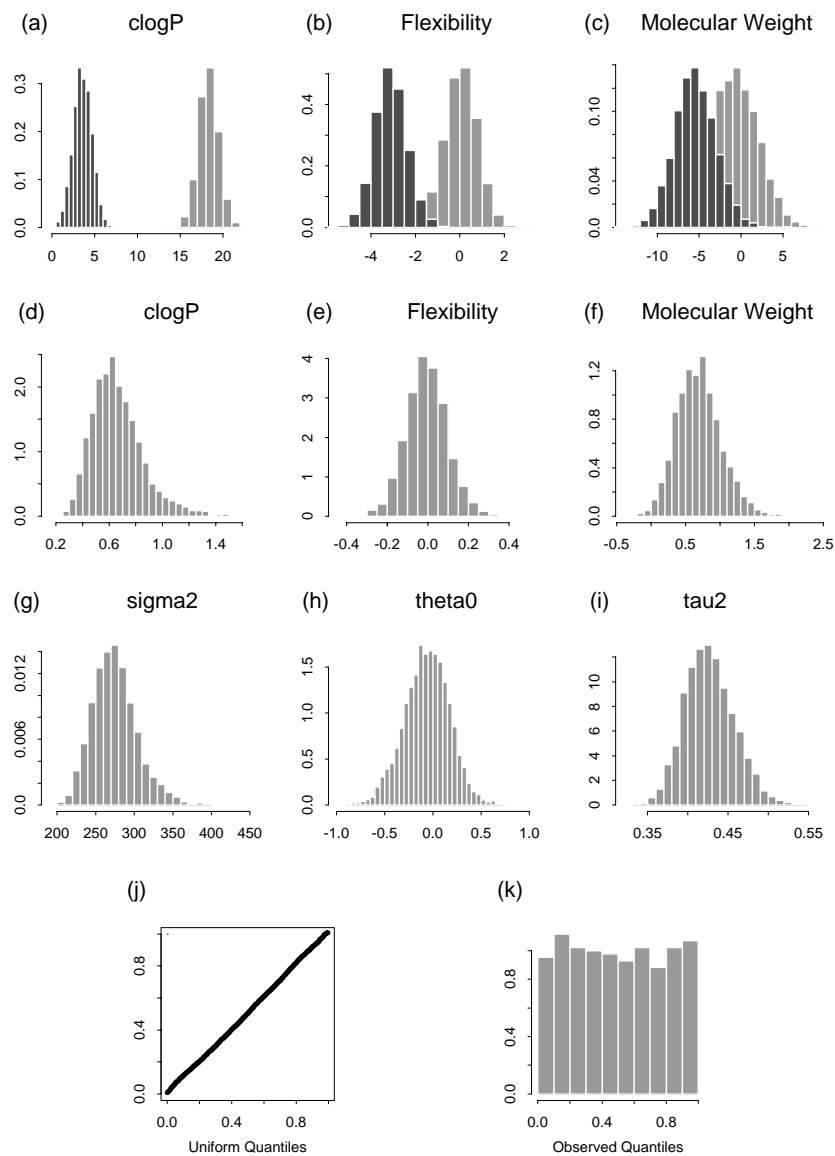


FIGURE 5. Posteriors in analysis of Glaxo Wellcome dataset. (a-c): β parameters in potency regressions. The right-most histograms are for regression coefficients for monomers binding to S_1 , and the left-most for monomers binding to S_2 . (d-f,h): Probit regression parameters θ . (g): σ^2 . (i): Binding probability variance parameter τ^2 . (j,k): Predictive quantiles of observed data in uniform quantile plot and histogram.

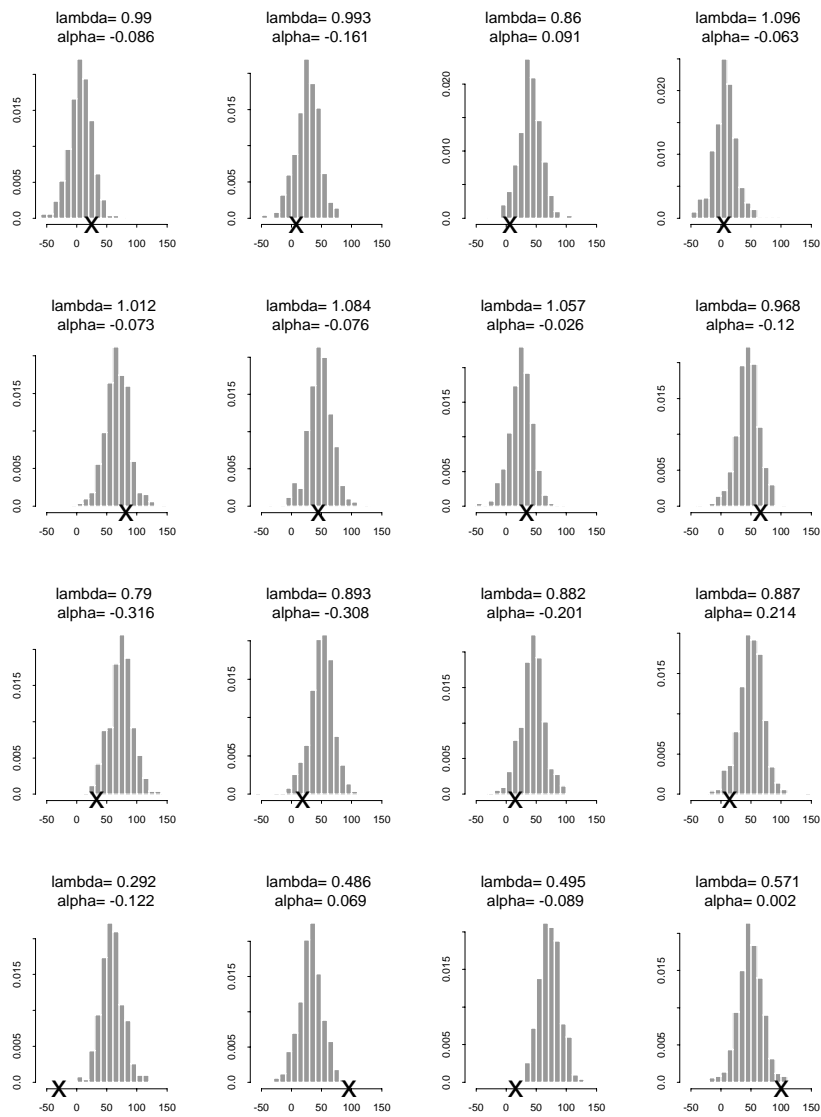


FIGURE 6. Predictive densities of eight randomly-selected compounds (rows 1 and 2), four compounds with extreme α_i values (row 3), and four compounds with extreme λ_i value (row 4). Data values appear as crosses on the axes.

to enhance computation of chemical descriptors of monomers, so providing additional covariates for analysis. Evidently, the real data analysis here isolates some meaningful relationships between covariates and both responses – binding and potency – but much more is needed to improve predictive ability for future compounds. Our discussion indicates that, with the extensions to incorporate hierarchical model components describing heterogeneity in both the binding/orientation selection mechanism and the potency outcome distributions, our model appears to adequately represent this particular real data set, in spite of a relatively high degree of assay and experimental variability. Posterior distributions on model parameters indicate subsets of the chosen covariates are informative about the structure-activity relationship, and we have contributed to the chemistry in novel ways by isolating and estimating the differential effects of covariates on binding orientation and potency scales separately. The high levels of assay variability limit predictive accuracy, though our derived predictions are generally accurate. Our random effects models are critical in accounting for compound-specific characteristics that are not measured or understood, and this may prove useful in future studies of similar compounds whose monomers have identical covariate values but, when analyzed, apparently distinct random effects. The framework may also be modified to include random effects for the monomers, rather than just the overall compounds, in support of such experimental developments.

Current research on the statistical side includes exploration of the identifiability problem, especially when the number of covariates is small and/or when the assay variance is large. We have also begun the development of related models for more complex experiments involving more than two monomers and receptor binding sites. Additionally, we need to explore more monomer covariates, perhaps many more than the simple few here. What has been learned in this simpler case is naturally guiding these challenging but potentially rewarding investigations.

Acknowledgments

This research is partially supported by Glaxo Wellcome, 5 Moore Drive, RTP, NC 27709, and the National Institute of Statistical Sciences (NISS), <http://www.niss.org>. Corresponding author is Susan Paddock, Institute of Statistics and Decision Sciences, Duke University, Durham, NC 27708-0251, USA, <http://www.stat.duke.edu>

References

- Albert, J.H. and Chib, S. (1993) Bayesian analysis of binary and polychotomous response data, *Journal of the American Statistical Association*, **88**, 669-679.
- Gelfand, A.E. and Smith, A.F.M. (1990) Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association*, **85**, 398-409.
- Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1996) *Bayesian Data Analysis*, London: Chapman and Hall.
- Titterton, D.M. and Smith, A.F.M., and Makov, U.E. (1985) *Statistical Analysis of Finite Mixture Distributions*, London: Wiley.
- Whitten, J.P., Xie, Y.F., Erickson, P.E., Webb, T.R., De Souza, E.B., Grigoriadis, D.E., and McCarthy, J.R. (1996) Rapid microscale synthesis, a new method of lead optimization using robotics and solution phase chemistry: Application to the synthesis and optimization of Corticotropin-releasing factor₁ receptor antagonists, *Journal of Medicinal Chemistry*, **39** 4354-4357.
- West, M. (1984) Outlier models and prior distributions in Bayesian linear regression. *Journal of the Royal Statistical Society (Ser. B)*, **46**, 431-439.
- West, M. (1997) Hierarchical mixture models in neurological transmission analysis. *Journal of the American Statistical Association*, **92**, 587-606.