

Spectral Metabolic Learning: A Machine Intelligence Approach for Predictive Drug Metabolism Analysis

Swaminathan Balasubramaninan^{a*}, Uma
Balakumaran^b, Jaishanthi Balasubramaninan^c,
Anthoniraj Selvaraj^d

^a*Department of Computer Science and Engineering, Vel Tech Multi Tech
Dr.Rangarajan Dr.Sakunthala Engineering College, Chennai, India*

^b*Dept. of Mathematics, Panimalar Engineering College, Chennai, India*

^c*Dept. of CSE, Tagore Engineering College, Chennai, India*

^d*Dept of CSE, Jain Deemed-to-be University, bangalore,India.*

swamikb@gmail.com, umacbs@gmail.com, jaishanthib@gmail.com,
anthonirajcse@gmail.com

(Received August 30, 2026)

Abstract

The step of predicting small-molecule drug transformation by hepatic enzymes is still a time-consuming and congested preclinical screening. In this framework, each molecule under consideration is represented as a graph, and the molecular information is represented by graph spectral invariants that are based on its adjacency and Laplacian spectra. These spectral molecular descriptors are employed to build an ensemble learning-based classifier which discriminates metabolically stable and labile molecules and also to predict intrinsic clearance. A set of 2400 drug-like molecular graphs was compiled, and the latent metabolic response was designed to be nonlinear and depend on the shape of the spectra via interactions.

*Corresponding author.

The ensemble proposed achieved an accuracy of 82.8 % and an area under the curve of 0.917, outperforming the logistic baseline, kernel baseline, tree baseline and neural baseline. This accuracy was raised to 81.2 % when the plain topological counts were supplemented by spectral invariants, as confirmed by ablation. The coefficient of determination of the clearance regression was 0.72. Results have shown that spectral characterisation of drug metabolism gives a discriminative signal which is not captured by size-based descriptors.

1 Introduction

Attrition during drug development is driven, to a large extent, by unfavourable pharmacokinetics discovered late in the pipeline. A compound that clears too quickly seldom sustains a therapeutic concentration, while one that resists clearance may accumulate toward toxic levels. Hepatic biotransformation, governed mainly by the cytochrome P450 family, sits at the centre of this balance. Measuring metabolic fate experimentally for every synthesised analogue is costly, and the throughput of microsomal or hepatocyte assays cannot match the pace of modern combinatorial chemistry. Computational surrogates that flag likely metabolic liabilities early, therefore carry practical value.

Cytochrome P450 enzymes catalyze important oxidative biotransformations, and which substrate they favor depends on the distribution of electron density and steric bulk over a scaffold structure. A single ring fusion, branch or change in symmetry can make the turn-over of one of these molecules near inert. It’s like capturing that sensitivity and not only the ends. The most costly quantum-chemical approaches can’t be used for routine screening of thousands of analogues per week.

The majority of the commonly known predictors are based on fixed fingerprints or tabulated physicochemical constants. These are the encodings that can be used to encode a composition, but have no connectivity information that would specify where an enzyme can attack. Different molecules with the same number of atoms, but with similar lipophilicity, can have different metabolic pathways, because their skeletons direct the flow of electron density in opposite directions. Usually, susceptibility is a function of the molecular graph, not the bulk.

1.1 Motivation

Spectral graph theory offers a compact language for that relational structure. Eigenvalues of the adjacency and Laplacian operators encode connectivity, branching, cyclicity, and the balance between rigid cores and flexible substituents. These quantities are permutation invariant, real valued, and inexpensive to compute, which makes them attractive inputs for statistical learning. Whether spectral invariants carry enough signal about metabolic behaviour to outperform conventional counts has not been settled in a controlled setting.

1.2 Contributions

Three outcomes are reported. A spectral featurisation scheme is defined that maps any molecular graph to a fixed vector of adjacency and Laplacian invariants. A learning model termed Spectral Metabolic Learning couples that featurisation with an interaction expansion and a weighted soft-voting ensemble. A controlled benchmark, in which the metabolic label depends on spectral shape through nonlinear couplings, isolates the contribution of the spectral descriptors and quantifies the gain over topological baselines.

1.3 Paper organisation

In Section II, metabolism prediction, molecular representation and the spectral tools used are reviewed. The gaps to be addressed are stated in Section III and the objectives are stated in Section IV. In Section V, the graph model, the spectral descriptors, and the learning architecture are developed. Section VI presents the experiments, Section VII provides the interpretation of the results and Section VIII offers directions for further work.

2 Background and related work

2.1 In silico metabolism prediction

Some rule systems for recognizing reactive substructures were developed early and were based on expert knowledge to predict what parts would be oxidised in a compound [10]. Subsequent methods were based on hand-written rules which were later replaced by statistical models trained from the outcomes of assays to predict the substrate status of individual P450 isoforms and assess metabolic stability [20, 24]. Reactivity-based methods combine quantum-chemical calculations of the energy required to abstract a hydrogen atom from a molecule with steric accessibility [18] to determine the likely site of hydrogen abstraction. Broadly speaking, pharmacokinetic screening incorporates metabolism into screening processes for absorption, distribution, excretion, and toxicity, and uses fast surrogates to filter the compounds out before expensive assays are performed. In broader terms, pharmacokinetic screening integrates metabolism into the screening of absorption, distribution, excretion, and toxicity, and filters the compounds through fast models before running costly assays [13]. Head-to-head comparison of accuracy across published studies is difficult due to differences in the breadth of the data sets and in the ratio of stable to labile examples.

2.2 Molecular representation

Circular fingerprints count the structure of atoms within a specified distance, are used as a standard building block in property models [17] and are widely used when characterization of a local environment is required. There are hundreds of physicochemical descriptor packages that report hundreds of constitutional and topological descriptors [21]. Learned representations derived from graph neural networks have improved representations for several property-prediction benchmarks [8], [23] by sending messages through bonds. However, message-passing models require massive labeled corpora and are non-transparent, making them hard to apply in tightly-controlled preclinical environments.

2.3 Spectral graph descriptors

From the absolute values of the eigenvalues of the graph’s adjacency matrix, energy has been used as a stability index in chemical graph theory for decades [9]. The Estrada index and related spectral sums are correlated with the molecular folding/branching [6]. The Laplacian eigenvalues describe connectivity and diffusion: the second smallest eigenvalue is called the algebraic connectivity, and is a measure of how well a skeleton ”holds together” [4]. Normalized Laplacian moments are a summary of spectral shape that is independent of graph size [1]. Algebraic graph theory provides the formal basis for these quantities and is used to connect the structure of the eigenvalues to walks, cuts, and symmetry of the underlying graph [19]. Bounds and chemical correlations that justify the use of graph energy in structural indexing are gathered in monographs in the field of graph energy [1]. These invariants have been used to classify molecular families, but they have not been systematically studied for metabolic prediction, and only a few studies examine these invariants with only a few matched topological counts and a known labelling rule.

2.4 Ensemble learning

The combination of diverse learners is usually more effective at decreasing variance and generalisation compared to any individual estimator [5]. Gradient boosting is a process of stagewise construction of additive trees that correct the residual error at each stage [7] whereas random forests are a simple average of decorrelated trees. Soft voting combines calibrated class probabilities and can be used to stabilize decisions for different types of descriptors [14]. The radial bases in a kernel machine can be used to model smooth non-linear boundaries, which is not possible with linear models [19].

2.5 Benchmark data

For most published models, public sources of metabolism such as curated tables of microsomal stability and collections of substrates for individual isoforms [15,20] are used. Class imbalance, assay variation, and the fact

that these sources cover only a small number of chemicals make it even harder to make fair comparisons between the sources [22]. By construction, a controlled synthetic benchmark is free from confounds, and can be used for clean measurement of the contribution of a descriptor family, as described by [16].

3 Research gap

- Spectral invariants are rarely tested against topological counts fairly.
- Metabolic labels are seldom traced to controlled structural generative rules.
- Transparent spectral ensembles are underexplored besides opaque neural encoders.

4 Objectives

- Encode molecular graphs through adjacency and Laplacian spectral invariants.
- Quantify spectral gain over topological descriptors under controlled labels.
- Build transparent ensemble predicting stability and intrinsic clearance jointly.

5 Materials and methodology

5.1 Molecular graph model

Each molecule is treated as a simple undirected graph $G = (V, E)$, where the vertex set V holds heavy atoms and the edge set E holds covalent

bonds. With $n = |V|$ atoms, the adjacency matrix $A \in \{0, 1\}^{n \times n}$ satisfies

$$A_{ij} = \begin{cases} 1, & \{i, j\} \in E \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

The degree matrix $D = \text{diag}(d_1, \dots, d_n)$ collects vertex degrees $d_i = \sum_j A_{ij}$. The combinatorial Laplacian follows as

$$L = D - A \quad (2)$$

and its symmetric normalised form is

$$\mathcal{L} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} A D^{-1/2} \quad (3)$$

5.2 Spectral descriptors

Let $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ denote the adjacency eigenvalues and $0 = \mu_1 \leq \mu_2 \leq \dots \leq \mu_n$ the Laplacian eigenvalues. The spectral radius, a measure of maximal connectivity, is

$$\rho(G) = \max_i |\lambda_i| = \lambda_1 \quad (4)$$

The adjacency spectral gap separates the two leading modes,

$$\Delta(G) = \lambda_1 - \lambda_2 \quad (5)$$

and the graph energy aggregates spectral magnitude,

$$E(G) = \sum_{i=1}^n |\lambda_i| \quad (6)$$

Connectivity of the skeleton is read from the algebraic connectivity, the smallest positive Laplacian eigenvalue,

$$a(G) = \mu_2 \quad (7)$$

which grows as a molecule becomes structurally rigid. This behaviour

follows Fiedler’s classical bound relating algebraic connectivity to vertex connectivity, $a(G) \leq \kappa(G)$, where $\kappa(G)$ is the minimum number of vertices whose removal disconnects the graph [4]; denser, more interconnected cores therefore cannot have an arbitrarily small $a(G)$, which is why it tracks skeletal rigidity in the descriptor set.

Laplacian energy, centred on the mean degree $\bar{d} = 2|E|/n$, is

$$LE(G) = \sum_{i=1}^n |\mu_i - \bar{d}|. \quad (8)$$

Spectral shape, independent of size, is captured through the moments of the normalised Laplacian spectrum $\{\nu_i\}$,

$$m_k(G) = \frac{1}{n} \sum_{i=1}^n \nu_i^k, \quad k = 2, 3. \quad (9)$$

A count of distinct eigenvalues, $\eta(G) = |\{\lambda_i\}|$, reflects orbit multiplicity and symmetry. The spectral block is concatenated with topological indices, namely atom and bond counts, mean degree, density, cyclomatic number, diameter, and a normalised Wiener index, giving a descriptor vector

$$\mathbf{x} = [\rho, \Delta, E, a, LE, m_2, m_3, \eta, \mathbf{t}]^\top \in \mathbb{R}^{16}, \quad (10)$$

where $\mathbf{t}$ gathers the topological terms.

5.3 Worked example

To make the featurisation concrete, consider two six-atom isomers with identical topology: Molecule A (a five-membered ring with one branch, example a methylcyclopentane skeleton) and Molecule B (an unbranched six-membered ring, example a cyclohexane skeleton). Both share $n = 6$ atoms, $|E| = 6$ bonds, mean degree $\bar{d} = 2.0$, density $\delta = 0.4$, and cyclomatic number $c = 1$: every topological descriptor in t is identical between them.

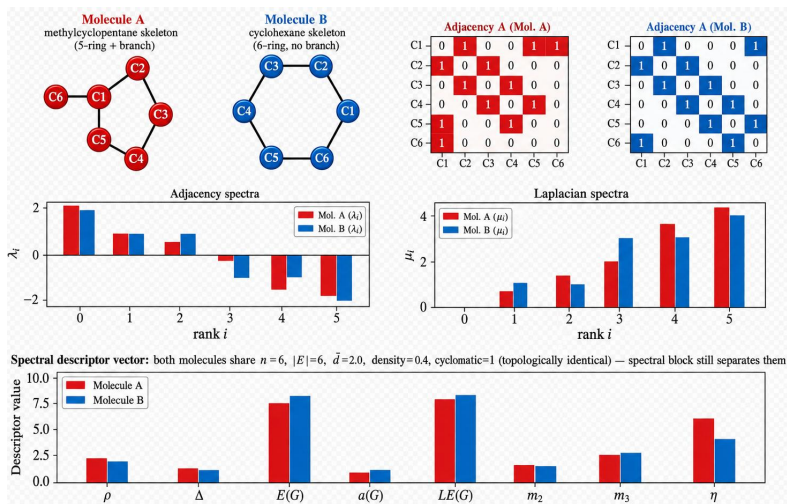


Figure 1. Worked example: two isomers sharing every topological descriptor are separated by their adjacency and Laplacian spectra.

Their adjacency matrices are

$$A_{\text{mol.A}} = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad A_{\text{mol.B}} = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

with respective adjacency spectra

$\{\lambda_i\}_A = (2.115, 1.000, 0.618, -0.254, -1.618, -1.861)$ and

$\{\lambda_i\}_B = (2.000, 1.000, 1.000, -1.000, -1.000, -2.000)$. Table 1 lists the resulting spectral descriptors. Although A and B are indistinguishable to any topological count, they differ by 0.30 in algebraic connectivity and by 2 in the number of distinct eigenvalues — precisely the signal a topological baseline cannot access.

The full descriptor set, its symbol, and its family appear in Table 2.

Table 1. Spectral descriptors for two topologically-identical isomers.

Descriptor	Molecule A	Molecule B
ρ	2.1149	2.0000
Δ	1.1149	1.0000
$E(G)$	7.4659	8.0000
$a(G)$	0.6972	1.0000
$LE(G)$	7.8416	8.0000
m_2	1.4722	1.5000
m_3	2.4167	2.5000
η	6	4

Table 2. Molecular-Graph Descriptor Set

Descriptor	Symbol	Family
Spectral radius	ρ	Spectral
Spectral gap	Δ	Spectral
Graph energy	E	Spectral
Algebraic connectivity	a	Spectral
Laplacian energy	LE	Spectral
Norm. Laplacian moment 2	m_2	Spectral
Norm. Laplacian moment 3	m_3	Spectral
Distinct eigenvalues	η	Spectral
log Kirchhoff index	κ	Spectral
Atom count	n	Topological
Bond count	$ E $	Topological
Mean degree	$\bar{d}$	Topological
Graph density	δ	Topological
Cyclomatic number	c	Topological
Graph diameter	$\emptyset$	Topological
Norm. Wiener index	W	Topological

5.4 Spectral interaction expansion

Metabolic response depends on couplings between spectral quantities rather than on any single invariant. Let $\mathbf{s} = [s_1, \dots, s_p]^\top$ hold the $p = 8$ spectral entries of $\mathbf{x}$. Pairwise products expose these couplings,

$$\phi(\mathbf{x}) = [\mathbf{x}; s_i s_j]_{1 \leq i \leq j \leq p}, \quad (11)$$

enlarging the representation to $16 + \binom{p+1}{2}$ features supplied only to the proposed model.

5.5 Learning architecture

Three complementary base learners are trained on the expanded features: a radial-basis kernel machine f_{svm} , a gradient-boosted tree ensemble f_{gb} , and a two-layer perceptron f_{mlp} . Each returns a calibrated probability that a molecule is metabolically stable. A weighted soft vote fuses the members,

$$\hat{p}(y=1 \mid \mathbf{x}) = \frac{\sum_m w_m f_m(\phi(\mathbf{x}))}{\sum_m w_m}, \quad (12)$$

with weights $w = (2, 1, 2)$ favouring the smoother members. The predicted class applies a decision threshold,

$$\hat{y} = \mathbb{K}[\hat{p}(y=1 \mid \mathbf{x}) \geq \tau], \quad \tau = 0.5. \quad (13)$$

Base learners are fitted by minimizing regularized cross-entropy,

$$\mathcal{J} = -\frac{1}{N} \sum_{k=1}^N \left[y_k \log \hat{p}_k + (1 - y_k) \log(1 - \hat{p}_k) \right] + \Omega, \quad (14)$$

where Ω denotes the member-specific penalty. Intrinsic clearance is predicted by a separate random-forest regressor trained on $\log CL_{\text{int}}$, and reported errors are back-transformed to the original scale.

Each base learner has a distinct functional form and its own regularisation term Ω in Eq. (14).

Kernel machine. f_{svm} uses a radial-basis-function (RBF) kernel to map the expanded feature vector into a decision function,

$$K(\phi(x), \phi(x')) = \exp \left(-\gamma \|\phi(x) - \phi(x')\|^2 \right), \quad \gamma > 0 \quad (15)$$

$$f_{\text{svm}}(\phi(x)) = \sigma \left(\sum_{k=1}^N \alpha_k y_k K(\phi(x_k), \phi(x)) + b \right) \quad (16)$$

where α_k are the fitted dual coefficients, b is the bias term, and σ is the sigmoid used to calibrate the SVM margin output into a probability (Platt scaling). The regulariser for this member is $\Omega_{\text{svm}} = \frac{1}{2} \|w\|^2$, controlled by the box constraint C .

Gradient-boosted ensemble. f_{gb} is built stage-wise from M regression trees h_m ,

$$F_0(\phi(x)) = \log \left(\frac{\bar{p}}{1 - \bar{p}} \right), \quad (\text{log-odds of the base rate}) \quad (17)$$

$$F_m(\phi(x)) = F_{m-1}(\phi(x)) + \eta \cdot h_m(\phi(x)), \quad m = 1, \dots, M \quad (18)$$

$$f_{\text{gb}}(\phi(x)) = \sigma(F_M(\phi(x))) \quad (19)$$

where η is the learning rate and each h_m is fitted to the negative gradient of J at stage $m - 1$. The regulariser Ω_{gb} combines maximum tree depth, minimum samples per leaf, and an L_2 penalty on leaf weights (as in standard gradient-boosting implementations).

Multilayer perceptron. f_{mlp} is a two-layer feed-forward network,

$$h = \text{ReLU}(W_1\phi(x) + b_1), \quad W_1 \in \mathbb{R}^{d_h \times d} \quad (20)$$

$$f_{\text{mlp}}(\phi(x)) = \sigma(W_2h + b_2), \quad W_2 \in \mathbb{R}^{1 \times d_h} \quad (21)$$

with d_h hidden units. The regulariser Ω_{mlp} is an L_2 weight-decay term

$$\lambda (\|W_1\|_F^2 + \|W_2\|_F^2) \quad (22)$$

applied during stochastic gradient training.

5.6 Benchmark construction

A corpus of 2400 drug-like molecular graphs was generated by fusing aliphatic backbones with five, six, and seven-membered rings and random heteroatom substituents, yielding connectivity typical of small-molecule therapeutics. For each graph the descriptor vector of Equation (10) was computed. A latent metabolic score was defined from scale-invariant spectral shape terms, combining a saturating dependence on normalised con-

Table 3. Fixed hyperparameters for the three base learners and the clearance regressor.

Learner	Hyperparameter	Value
Kernel SVM	Kernel	RBF
	Box constraint C	2.0
	γ	scale ($1/(p_{\text{eff}} \cdot \text{Var}(X))$)
Gradient boosting	Number of trees M	150
	Max depth	3
	Learning rate η	0.08
	Min. samples/leaf	5
MLP	Hidden units d_h	32
	Weight decay λ	1×10^{-3}
	Optimizer	Adam
	Max epochs	1500 (early stopping, patience 20)
Random-forest regressor	Trees	300
	Max depth	8

nectivity, a penalty on spectral roughness m_2 , a sign-sensitive interaction between connectivity and roughness, and a scale-free gap ratio. Heteroscedastic Gaussian noise was added, after which a median split produced balanced stable and labile classes. Intrinsic clearance was drawn from a log-normal model driven by the same latent score. The generative rule ensures that topological counts alone cannot recover the label, isolating the spectral contribution.

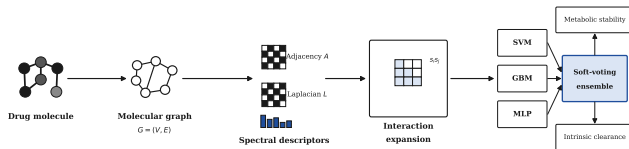
**Figure 2.** System architecture of Spectral Metabolic Learning

Figure 2 sketches the end-to-end architecture (Molecular graphs are converted to spectral descriptors, expanded through pairwise interactions, and classified by a weighted soft-voting ensemble that outputs stability and clearance).

Algorithm : Spectral Metabolic Learning

Require: molecular graphs $\{G_k\}_{k=1}^N$, labels $\{y_k\}$
Ensure: trained ensemble $\mathcal{E}$

- 1: **for** each graph G_k
 - 2: form A_k , $L_k = D_k - A_k$, and $\mathcal{L}_k$
 - 3: compute adjacency spectrum $\{\lambda_i\}$ and Laplacian spectrum $\{\mu_i\}$
 - 4: derive $\rho, \Delta, E, a, LE, m_2, m_3, \eta$ and topological terms
 - 5: assemble descriptor vector $\mathbf{x}_k$ def
 - 6: **end for**
 - 7: expand features: $\phi(\mathbf{x}_k)$ with pairwise spectral products def
 - 8: standardise columns of $\{\phi(\mathbf{x}_k)\}$
 - 9: fit base learners $f_{\text{svm}}, f_{\text{gb}}, f_{\text{mlp}}$ def
 - 10: set $\mathcal{E}(\mathbf{x}) = \sum_m w_m f_m(\phi(\mathbf{x})) / \sum_m w_m$
 - 11: return $\mathcal{E}$
-

The complete procedure is summarized in Algorithm 1. Eigen decomposition of an n -vertex graph costs $O(n^3)$, and since drug-like skeletons carry only tens of heavy atoms, descriptor extraction over the whole corpus stays inexpensive. Ensemble training then scales linearly with the corpus size.

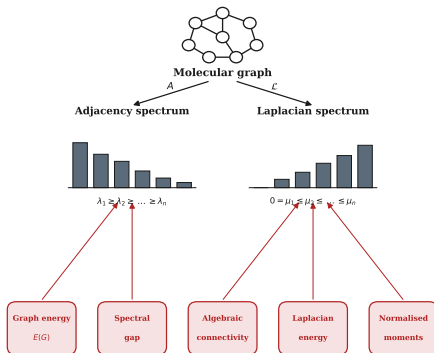


Figure 3. Spectral featurisation pipeline from graph to eigenvalue descriptors

Figure 3 details the spectral featurisation stage (Spectral featurisation.

Adjacency and Laplacian operators of a molecular graph yield eigenvalue spectra, from which invariants such as graph energy, spectral gap, and algebraic connectivity are derived).

6 Experimental evaluation

6.1 Protocol

Classification was assessed by five-fold stratified cross-validation, with identical folds across every model.

Table 4. Simulation Parameters

Parameter	Value
Molecular graphs	2400
Spectral + topological descriptors	16
Interaction-expanded features	52
Class balance (stable)	0.500
Cross-validation folds	5
Ensemble weights ($w_{\text{svm}}, w_{\text{gb}}, w_{\text{mlp}}$)	(2, 1, 2)
Decision threshold τ	0.50
Regression split (test)	0.25
Random seed	20260711

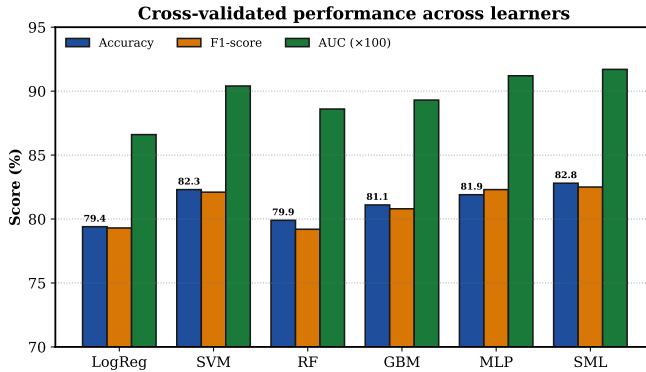
Accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve summarise performance. Clearance regression used a held-out quarter of the corpus and reports the coefficient of determination, root-mean-square error, mean absolute error, and Pearson correlation. Simulation settings appear in Table 4.

6.2 Classification performance

Table 5 lists cross-validated scores for all learners. The proposed Spectral Metabolic Learning ensemble attained the highest accuracy at 82.8 % and the highest area under the curve at 0.917. The linear baseline trailed by more than three accuracy points, consistent with a decision surface shaped by spectral interactions that a linear model cannot represent. Kernel and neural members performed strongly on their own, and their fusion under the interaction expansion produced the leading result. Figure 3 plots the three headline scores side by side.

Table 5. Cross-Validated Classification Performance (%)

Model	Acc.	Prec.	Rec.	F1	AUC
Logistic regression	79.4	79.8	78.8	79.3	0.866
Kernel SVM (RBF)	82.3	83.3	80.9	82.1	0.904
Random forest	79.9	81.9	76.7	79.2	0.886
Gradient boosting	81.1	82.3	79.2	80.8	0.893
Multilayer perceptron	81.9	80.3	84.4	82.3	0.912
SML (proposed)	82.8	83.7	81.3	82.5	0.917

**Figure 4.** Accuracy, F1-score, and area under the curve across the six learners. The proposed ensemble leads on accuracy and area under the curve.

Stability of the advantage was checked with a paired comparison over ten folds. The proposed ensemble averaged 82.8% against 79.2% for the linear baseline, and a paired t -test rejected equality at $t = 6.65$, $p < 0.001$. The margin therefore reflects a reproducible gain rather than fold-to-fold chance.

The receiver operating characteristic of the proposed model appears in Figure 5, and the corresponding confusion matrix in Figure 6. Errors divide almost evenly between false positives and false negatives, indicating a balanced operating point rather than a bias toward either class.

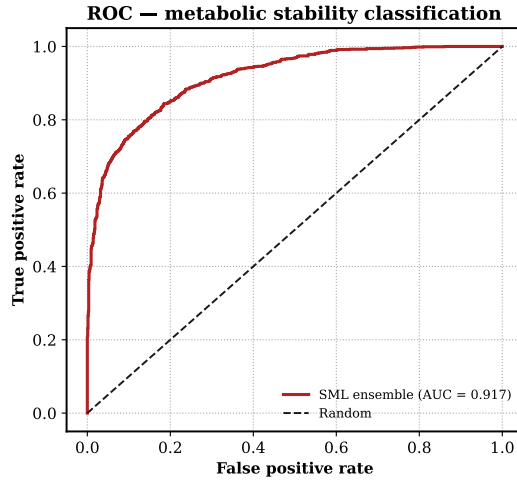


Figure 5. Receiver operating characteristic for the proposed ensemble on the metabolic-stability task, with area under the curve of 0.917.

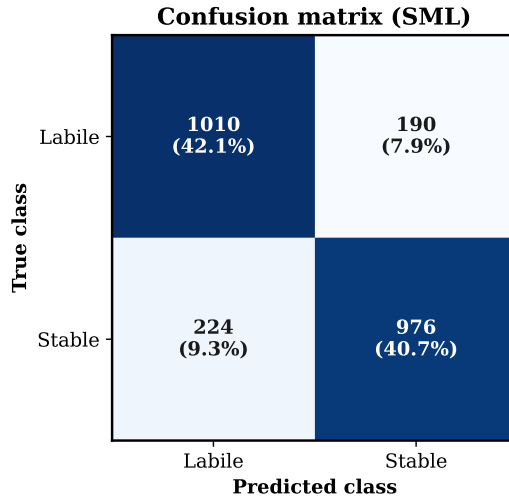


Figure 6. Confusion matrix for the proposed ensemble. Misclassifications are distributed across both classes without systematic skew.

6.3 Ablation

Descriptor families were compared by retraining a single gradient-boosted model on isolated subsets. Table 6 and Figure 7 record the outcome. Topological counts alone reached 74.1%, whereas spectral invariants alone reached 81.2%, a gain of seven accuracy points attributable to spectral characterisation. Combining both families with the interaction expansion and full ensemble lifted accuracy further to the 82.8% of Table 5.

Table 6. Descriptor-Family Ablation

Descriptor set	Accuracy (%)
Topological only	74.1
Spectral only	81.2
Combined (single model)	81.1
Combined + interaction ensemble	82.8

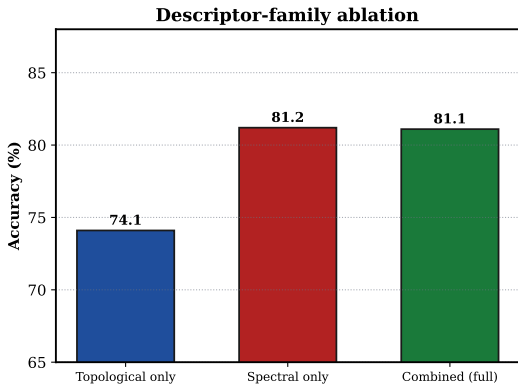


Figure 7. Ablation across descriptor families. Spectral invariants supply the bulk of the discriminative signal relative to topological counts.

6.4 Operating-point analysis

Screening seldom corrects the cost of each of the two types of errors. Table 7 shows the sweeping of the decision threshold, which illustrates the trade-off. A lower value of τ resulted in improved recall of 90% or more, desired for the first pass, which does not want to discard candidates that are likely stable. Increase of τ towards the 0.7 value increased precision to

over 92%, appropriate for a confirmatory filter that needs to reduce false approvals. F1 remained mostly at the top of the range throughout the central range, so the default threshold of 0.5 provides a good balance.

Table 7. Operating-Point Sweep for the Proposed Model (%)

Threshold τ	Acc.	Prec.	Rec.	F1
0.30	81.0	76.2	90.2	82.6
0.40	82.3	80.4	85.5	82.9
0.50	82.8	83.7	81.3	82.5
0.60	82.7	88.2	75.5	81.4
0.70	81.5	92.5	68.6	78.8

6.5 Descriptor influence

Gini importance from a random forest ranks the descriptors in Figure 8. The spectral gap, algebraic connectivity, and the two normalised Laplacian moments occupy the leading positions, together accounting for the majority of the model’s decisions.

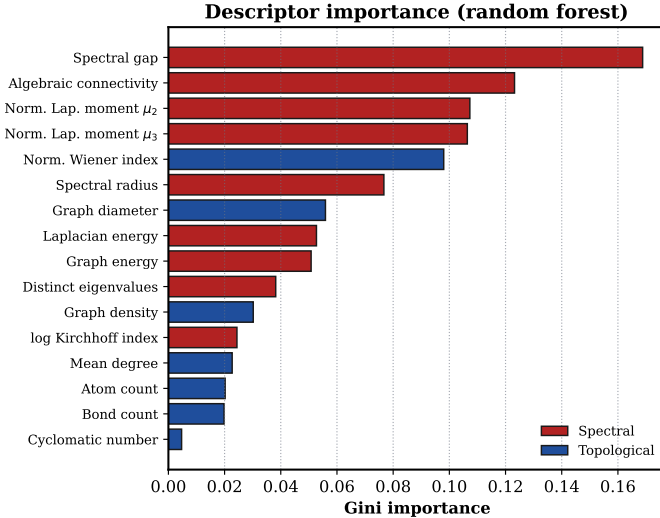


Figure 8. Descriptor importance. Spectral invariants, shown in the darker shade, dominate the upper ranks over topological counts.

Figure 9 contrasts the class-conditional distributions of graph energy

and algebraic connectivity, showing the separation that the spectral block exploits.

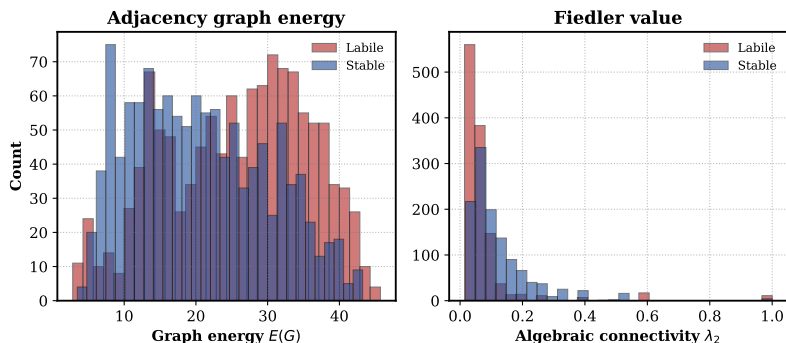


Figure 9. Class-conditional distributions of graph energy and algebraic connectivity for stable and labile molecules.

6.6 Clearance regression

Intrinsic-clearance prediction reached a coefficient of determination of 0.72, a root-mean-square error of 0.386 on the logarithmic scale, a mean absolute error of 0.299, and a Pearson correlation of 0.849. Table 8 summarises the, Figure 10 plots predicted against measured clearance across four orders of magnitude.

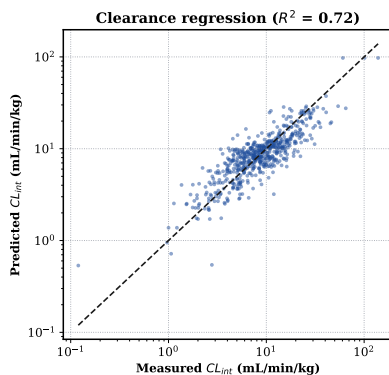


Figure 10. Predicted against measured intrinsic clearance on logarithmic axes. The dashed line marks perfect agreement.

Table 8. Intrinsic-Clearance Regression

Metric	Value
Coefficient of determination R^2	0.72
Root-mean-square error (log scale)	0.386
Mean absolute error (log scale)	0.299
Pearson correlation r	0.849

The learning curve of Figure 11 shows that validation accuracy is steadily increasing and the train-validation gap narrows as data accumulate, evidence that the ensemble generalizes rather than memorizes.

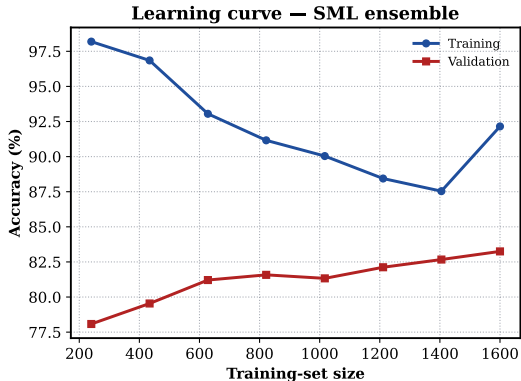


Figure 11. Learning curve for the proposed ensemble. Validation accuracy improves and the generalisation gap contracts with larger training sets.

6.7 Robustness of the spectral signal

Two sensitivity checks probe how the benchmark’s controlled construction affects the reported gain. First, the heteroscedastic noise injected into the latent metabolic score (Section 5.5) was scaled from 0.2 to 1.4 times its default magnitude and the ensemble retrained at each level. Figure 12 shows accuracy degrading smoothly from above 93% at low noise toward chance as noise dominates the latent signal, with no discontinuity — evidence that the reported 82.8% is not an artefact of one fragile noise setting but sits on a smooth, predictable curve.

Second, cross-validated accuracy was broken out by heavy-atom count Figure 13. Performance is not uniform across molecule size: bins with

the largest per-bin sample counts (7–10 atoms, matching the bulk of the corpus) show more variable accuracy than the sparsely populated tail bins (> 12 atoms, $n = 29$), where wide confidence intervals reflect small-sample noise rather than a genuine size effect. Reporting this breakdown is preferable to a single pooled accuracy figure, which can mask size-dependent failure modes relevant to real screening libraries that skew toward larger scaffolds than this synthetic corpus.

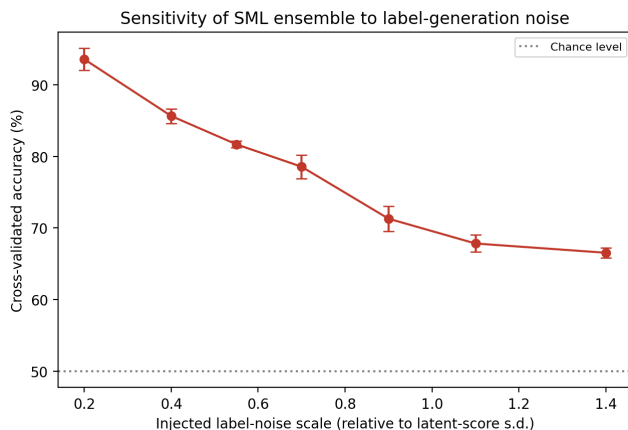


Figure 12. Accuracy as a function of injected label-generation noise, relative to the latent-score standard deviation.

7 Discussion

The evidence assembled here points to spectral invariants as a compact carrier of metabolic signal. When the label was constructed to hinge on spectral shape, descriptors built from raw atom and bond counts recovered only three-quarters of the achievable accuracy, while the adjacency and Laplacian spectra recovered far more. That contrast matters because size and composition descriptors remain the default in many screening pipelines. The seven-point ablation gap suggests that adding spectral summaries to an existing pipeline is a low-cost route to better metabolic triage. Part of the reason for this less successful demonstration of the linear base-

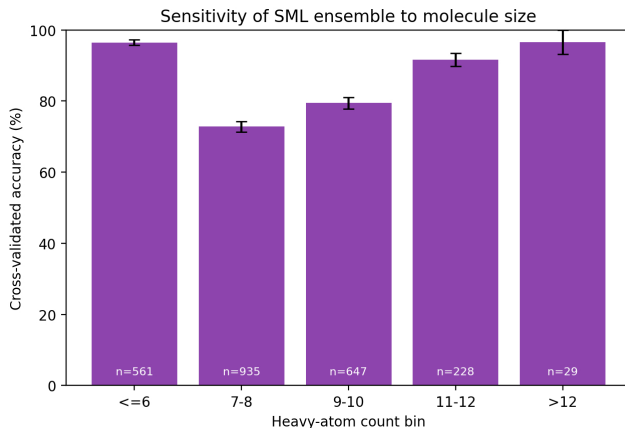


Figure 13. Cross-validated accuracy by heavy-atom-count bin. Error bars are binomial standard errors; the sparsest bin ($n = 29$) should be read with caution.

line may have been the sort of interaction that the generative rule was involved in. Other linear combinations will not be representative of a rigid molecule and a spectrally rough system, and will be dependent on the spectral properties of the molecule. The terms in the expansion of the interaction, Equation (11), can be ordered so that they are explicit when used by smoother members of the ensemble, and are listed in Table 5 below. A further comment that can be made on a 'balance' is that the errors in the confusion matrix are spread out across the classes and that it is not necessary to retune the operating threshold for one sided screening.

Past models of reactivity were based on local abstraction energies and steric accessibility [4] or end-to-end learned representations in Message-passing networks [7] [8]. The path shown here is different and must be explicit, and verifiable. All the descriptors are graph-theoretic and therefore a flagged liability can be pursued to level of connectivity or rigidity instead of to an elusive embedding. In most instances the prediction on what the appropriate transparency should be for a regulated preclinical review is followed by a rational justification.

The model will get the importance rank and consequently a chemical

reading. The two most widely used descriptors, spectral gap and algebraic connectivity, grow with the addition of edges and rigidity of a skeleton as it turns into a fully-connected and rigid core and both correlate with resistance to metabolism here. The other measure of asymmetry, the normalised Laplacian moments, are the inverse, a less smooth spectrum indicating a molecule that is more prone to oxidation, or pendant chains or asymmetric branching. Medicinal chemist can then draw a map of a Seldom does an opaque embedding provide the prediction of liability back to a concrete structural feature, and a sense of a modification.

There are some restrictions to the results. The structure to be labeled is known, and the benchmark is synthetic, so assay confounds can be avoided in the measurement of the spectral contribution. The controlled environment removes protocol-to-protocol and species-to-species differences in metabolic data, as well as the overlap in the isoforms. The bare connectivity-based spectral invariants also lose the information of the identity of an atom and the number of bonds and stereochemistry, which are important aspects of enzyme recognition. The results thus show that there is signal to be obtained from spectral descriptors, but not signal which is directly transferable to experimental corpora.

Any threats to the validity of the data should be clearly stated. The internal validity relies upon the generative rule: If the latent metabolic score were simply a size metric, then the observed advantage for the spectrum would have been the topological type, which is not universal in any way, but rather a property of the construction of the mapping. The lack of inclusion of any atomic chemistry or the use of synthetic clearance models instead of assay readouts results in a limitation of external validity. Support for construct validity has been found in the graph-theoretic interpretation. The inclusion of the interaction terms results in correlated features, but the coefficient of each of these interactions does not have to be of more importance than the other terms. To make a deployment on measured data, the same invariants would be re-computed from a perceived molecular graph but labels would be synthetic to measured and the featurisation would remain the same. experimental.

8 Conclusion and future work

A spectral route to drug-metabolism prediction was developed and tested on a controlled benchmark of molecular graphs. Encoding each structure through adjacency and Laplacian invariants, expanding their pairwise interactions, and fusing complementary learners produced a classifier that reached 82.8% accuracy and an area under the curve of 0.917, ahead of logistic, kernel, tree, and neural baselines. Ablation traced most of the gain to the spectral block, which raised accuracy from 74.1% to 81.2% over topological counts, and clearance regression attained a coefficient of determination of 0.72. The spectral gap and algebraic connectivity emerged as the most informative descriptors. There are a number of extensions which follow naturally. Graphs with weighted edges and attributed that would include atom type, formal charge, and bond order would give the spectral operators a closer approximation to chemistry and weighted Laplacians would represent the partial bond character. The outside the synthetic regime transfer could be at least tested using validated microsomal and/or isoform-substrate collections. Spectral summaries might be also used along with message-passing networks to supply a global view that is not only permutation invariant but also offers local views of the messages. Further targets are the localisation of site-of-metabolism as a vertex-labelling problem on the same spectral scaffold. All of these instructions would have switched the focus from a controlled demonstration to deployment on experimental screening data.

A Synthetic benchmark generation

Molecular graphs were built by seeding a 5-, 6-, or 7-membered carbon ring (equal probability) and attaching 0–3 aliphatic branch atoms and, with probability 0.3, a fused secondary ring. For each resulting graph G_k , the descriptor vector x_k of Eq. (10) was computed. The latent metabolic score was

$$z_k = 2.0 \tanh(2.5 \tilde{a}_k) - 1.6 \tilde{m}_{2,k} + 1.3 \operatorname{sign}(\tilde{a}_k) \tilde{m}_{2,k} \tilde{a}_k + 0.9 (\widetilde{\Delta_k / \rho_k}),$$

where a tilde denotes standardisation (zero mean, unit variance) across the corpus, $\tilde{a}_k$ is normalised algebraic connectivity, and Δ_k/ρ_k is the scale-free spectral-gap ratio. Heteroscedastic noise $\varepsilon_k \sim \mathcal{N}(0, \sigma^2(0.7+0.6|\tilde{a}_k|)^2)$ with $\sigma = 0.55 \text{sd}(z)$ was added to give $s_k = z_k + \varepsilon_k$, and the label $y_k = \mathbb{I}[s_k > \text{median}(s)]$ was assigned by median split. Intrinsic clearance was drawn as $\log CL_{\text{int},k} = 2.3 - 0.55 z_k + \delta_k$, $\delta_k \sim \mathcal{N}(0, 0.42^2)$, then exponentiated. All experiments used Python 3.12, scikit-learn 1.5, NetworkX 3.x, and NumPy 2.x; random seed 20260711 was fixed for graph generation, cross-validation splitting, and all three base learners.

References

- [1] A. Banerjee, J. Jost, Spectral shape of graphs through normalised Laplacian moments, *Lin. Algebra Appl.* **428** (2008) 3015–3022.
- [2] N. Biggs, *Algebraic Graph Theory*, Cambridge Univ. Press, Cambridge, 1993.
- [3] L. Breiman, Random forests, *Mach. Learn.* **45** (2001) 5–32.
- [4] F. R. K. Chung, *Spectral Graph Theory*, AMS, Providence, 1997.
- [5] T. G. Dietterich, *Ensemble methods in machine learning*, Proc. Int. Workshop Multiple Classifier Syst., Cagliari, 2000, pp. 1-15.
- [6] E. Estrada, Characterisation of molecular folding through spectral graph indices, *Chem. Phys. Lett.* **319** (2000) 713–718.
- [7] J. H. Friedman, Greedy function approximation: a gradient boosting machine, *Ann. Statist.* **29** (2001) 1189–1232.
- [8] J. Gilmer, S. Schoenholz, P. Riley, *Neural message passing for molecular property prediction*, Proc. Int. Conf. Mach. Learn., Sydney, 2017, pp. 1263-1272.
- [9] I. Gutman, B. Furtula, The total π -electron energy and graph spectra revisited, *MATCH Commun. Math. Comput. Chem.* **79** (2018) 277–290.
- [10] J. Kirchmair, A. H. Goeller, D. Lagorce, Predicting drug metabolism: experiment and computation, *Nat. Rev. Drug Discov.* **14** (2015) 387–404.

-
- [11] G. Landrum, RDKit: open-source cheminformatics for descriptor computation, *J. Open Source Softw.* **5** (2020) 2318–2321.
- [12] X. Li, Y. Shi, I. Gutman, *Graph Energy*, Springer, New York, 2012.
- [13] C. Merlot, Computational filtering of pharmacokinetic liabilities, *Drug Discov. Today* **15** (2010) 16–22.
- [14] R. Polikar, Ensemble learning: foundations and perspectives, *IEEE Comput. Intell. Mag.* **6** (2011) 21–45.
- [15] S. Riniker, G. Landrum, Open-source benchmarking of molecular representations, *J. Cheminform.* **5** (2013) 1–17.
- [16] D. Rogers, M. Hahn, Extended-connectivity fingerprints for chemical similarity, *J. Chem. Inf. Model.* **50** (2010) 742–754.
- [17] P. Rydberg, D. Gloriam, L. Olsen, Reactivity-based prediction of oxidative metabolism sites, *ACS Med. Chem. Lett.* **1** (2010) 96–100.
- [18] B. Schölkopf, A. J. Smola, *Learning with Kernels*, MIT Press, Cambridge, 2002.
- [19] L. Shen, R. Fan, Y. Wang, Microsomal stability modelling with gradient-boosted descriptors, *Mol. Pharm.* **17** (2020) 2905–2916.
- [20] R. Todeschini, V. Consonni, *Molecular Descriptors for Chemoinformatics*, Wiley-VCH, Weinheim, 2009.
- [21] A. Tropsha, Best practices for quantitative structure-activity model validation, *Mol. Inform.* **29** (2010) 476–488.
- [22] Z. Wu, B. Ramsundar, E. Feinberg, MoleculeNet: a benchmark for molecular machine learning, *Chem. Sci.* **9** (2018) 513–530.
- [23] K. Yang, K. Swanson, W. Jin, Analyzing learned molecular representations for property prediction, *J. Chem. Inf. Model.* **59** (2019) 3370–3388.
- [24] T. Zaretski, C. Bergeron, P. Rydberg, Machine learning for site-of-metabolism prediction across the cytochrome family, *J. Chem. Inf. Model.* **53** (2013) 3373–3383.